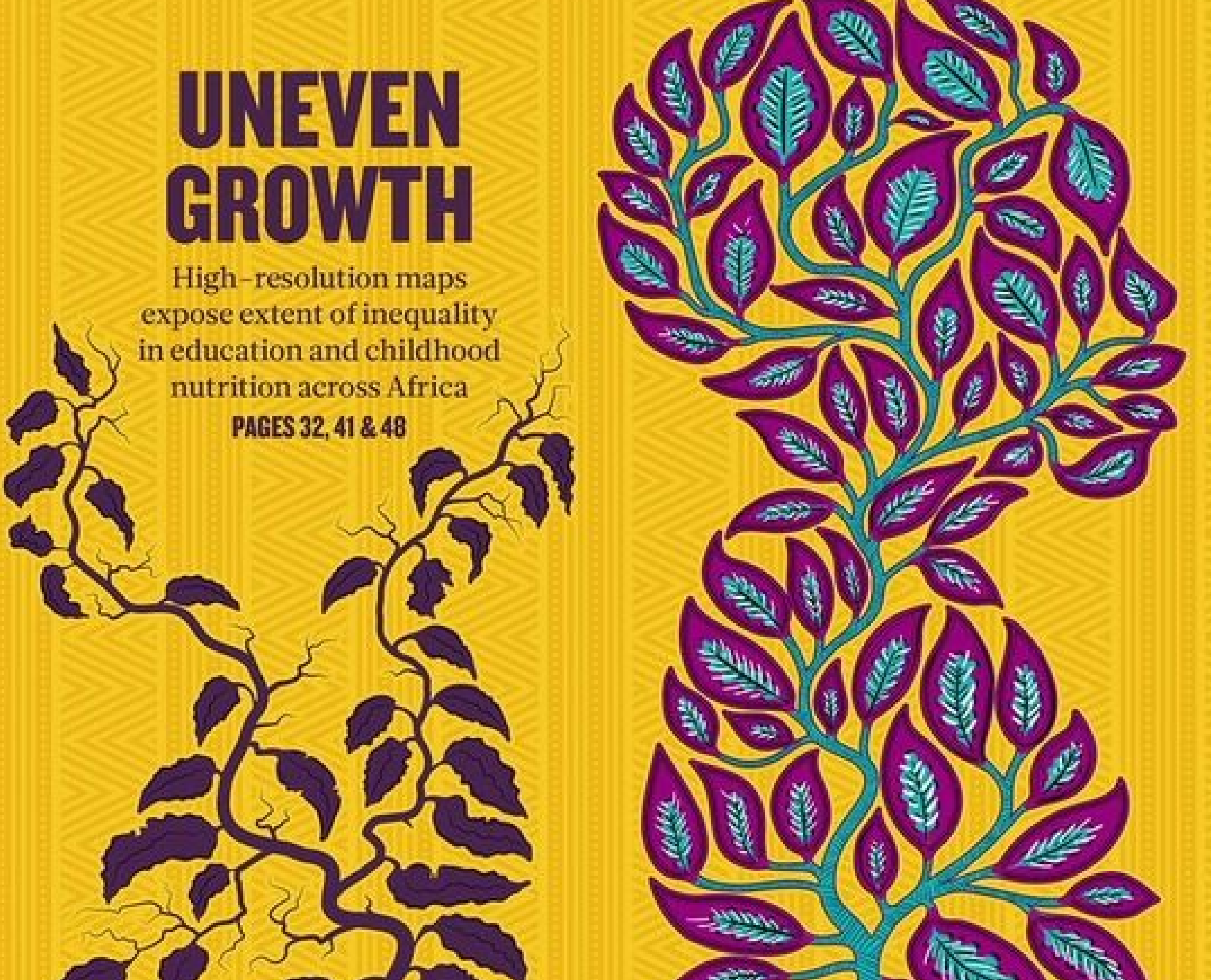
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

UNEVEN GROWTH

High-resolution maps expose extent of inequality in education and childhood nutrition across Africa

PAGES 32, 41 & 48



NATURE.COM/NATURE

1 March 2018 · $10

Vol. 555, No. 7694

THIS WEEK

EDITORIALS

REPRODUCIBILITY Reporting summaries to promote transparency **p.6**

WORLD VIEW Invest in data science to tackle malnutrition **p.7**

WING SPOTS Gene identified that produces double-sided design **p.9**

How to be good

A proposed international code of ethics can serve as a springboard for discussion on how scientists can take positive action in their own workplace.

You are an early-career scientist poised to publish a paper that you think will be your big break. It describes your imaginative hypothesis — a potential scientific insight with substantial implications — along with the experiments you designed and constructed, and the carefully documented data that support your initial insight. It's a genuine advance for the field and will be widely cited. Your lab head will be satisfied. Job done!

Then, disaster. You wake in the small hours and realize a possible flaw: another way in which the data could be interpreted, that would throw the conclusion into doubt. No one else will spot the problem — the lab head is too busy and no editor or reviewer will realize — and further experiments to settle the issue will take time. Worse, fresh results could sink the hypothesis (and subsequent grants). So, do you publish anyway?

Of course not! Science puts the pursuit of truth above all else, right? Well, not always. The dilemma above is a real one faced by real scientists, and not all of them jump the right way. What can help them to make the right decision? Some scientists think it might help to discuss this idea: "Pursuing the truth means following the research where it leads, rather than confirming an already formed opinion."

That statement opens one of seven presentations in a 'Code of Ethics for Researchers' produced by a group of scientists convened by the World Economic Forum. These scientists, drawn from many countries, are all under 40 but well established in career terms, with decades of research and leadership ahead of them. This combination makes them well qualified to explore the realities and pressures of modern lab life, so their ideas deserve to be considered by the scientific community.

Many science organizations have issued similar recommendations to their own research communities — the Science Council of Japan, for example, has made a valiant effort. But it can be difficult to persuade busy and pressured scientists to take notice of such guidelines, especially when — usually — they are expressed in rather terse form, as if on tablets of stone. This document, carrying the weight of an international consensus, demonstrates well how consideration of ethical issues is not additional to research practice, but an integral and essential component. It has the virtue of being presented in an engaging and persuasive style.

Each of the seven pages is headed by an exhortation such as "Minimize harm" and "Support diversity", backed by an explanation of why the entreaty matters and a brief presentation of the ultimate goal and practical approaches to achieving it. A collection of real-life anecdotes helps to illustrate their relevance.

The effort is valuable because, alongside fundamental tenets of research behaviour (such as "Be accountable" and "Be a mentor"), the code contextualizes natural sciences in a time of rapid technological change and popular questioning of expertise. Its authors see it as helping to redefine "the social and moral contracts that bind researchers to society" and infuse research with "the most irreproachable behaviours".

Some of this context is familiar: it is hardly original to seek to minimize harm to citizens (ranging from wasted public money to damage to health and the environment), or to say that engaging with the public is important. But there are more radical suggestions, too: that such engagement should sometimes include public involvement in the "identification of the question, conception of a project, discussion of results and dissemination". Some will see that as extreme. Yet many research projects fail to make the societal impacts they aim for because they lack precisely this depth and breadth of engagement.

Any well informed reader will spot ways in which this code is contentious to researchers, and could find ways to fault it. But that would be to miss its virtues as a focus for discussion, not just by active researchers but also by those in positions of influence in universities, research institutions, governments and private funding bodies. Nothing in the document is fundamentally new, and yet it will still be interpreted by many as highly aspirational and even unrealistic. Who in the real world, critics might scoff, would be willing to divert funding from postdoctoral posts into better training for principal investigators or other means by which the code can be better pursued?

"Consideration of ethical issues is not additional to research practice, but an integral and essential component."

As the authors state, their purpose is to stimulate open conversations "to safeguard a positive and sound research environment". Accordingly, *Nature* readers may do themselves and others some good by visiting http://wef.ch/coe and providing feedback. Even better, they might discuss the ideals expressed, and consider how to live up to them in their own lab, research institution or funding agency. We at *Nature* are trying to do so, too. ■

Wasted energy

Oil and gas emissions could exceed current estimates — and governments need to act.

When it comes to harmful emissions from the oil and gas industry, much of the focus tends to fall on methane, a potent greenhouse gas. But ethane and propane deserve scrutiny too.

Scientists know that ethane emissions rose during the early twentieth century and then began to decline in the 1970s as modern air regulations took effect. More recent monitoring suggests that the trend reversed in 2009, when atmospheric concentrations again began to rise, probably due to emissions from the expanding oil and gas industry, and particularly in the United States. Yet reconciling the atmospheric data with current emissions inventories has been tough.

This week in *Nature Geoscience*, researchers report progress (S. B. Dalsøren *et al. Nature Geosci.* http://dx.doi.org/10.1038/s41561-018-0073-0; 2018). The team simulated multiple emissions scenarios in an effort to reproduce observational data, including those gleaned from ice cores. They accounted for natural emissions from sources such as geological seeps and mud volcanoes, and plugged in detailed information about emissions from the fossil-fuel industry. The results suggest that the industry's ethane and propane emissions have been drastically underestimated, and are two to three times higher than figures used by the Intergovernmental Panel on Climate Change.

That in itself is a concern, because the two gases contribute to smog — but the study also underscores troublesome questions about industrial emissions of methane, which are second only to those of carbon dioxide when it comes to warming the planet. Where there are ethane and propane being released, there tends to be methane too. The ratios vary across oil and gas fields — but, once understood, they can be used to differentiate industrial methane emissions from those from other sources, including livestock, rice paddies and wetlands. As it happens, atmospheric methane concentrations also resumed their historical rise in 2007, after a nearly decade-long plateau.

This rise, some scientists have argued, implies that US methane emissions are also underestimated — they could be as much as double the US government's inventory — and that the fossil-fuel industry is largely to blame. But questions remain. Another line of research using carbon isotopes to determine methane's source found no such rise in industrial emissions, and suggested that the culprit for spiking methane concentrations is more likely to be rampant agricultural expansion in the tropics.

Governments don't need to know all the details before they act. The International Energy Agency estimates that the industry emits roughly 76 million tonnes of methane per year globally, and that three-quarters of those emissions could be eliminated with current technologies if companies fixed or replaced leaky equipment. Implementing just those measures that pay for themselves would be akin to reducing carbon dioxide emissions by 160 billion tonnes by 2100 — nearly 47 times the annual emissions of the European Union.

"Governments don't need to know all the details before they act."

Sadly, the administration of US President Donald Trump is moving in the opposite direction, undermining rules intended to reduce methane emissions. So far, courts have blocked these efforts, but the battle is heating up. On 22 February, the US Bureau of Land Management proposed a rule that would loosen requirements on oil and gas companies to reduce methane emissions.

The message is getting through internationally. Last year, a coalition involving industry, academia and environmentalists launched a research initiative to better document where and how much methane is being emitted from oil and gas operations. In November, eight of the world's major oil and gas producers committed to shoring up their operations. They understand the case for plugging leaks; now the United States must catch up. ■

ANNOUNCEMENT

Two documents for greater transparency

As part of a broader effort to improve reporting quality, *Nature* and the Nature journals introduced a reporting checklist for life-sciences papers in 2013. This asked authors to reveal some key details of experimental design. Last year, this checklist evolved into a broader reporting-summary document that is published alongside manuscripts to promote greater transparency.

We have now developed two new versions of the reporting summary: one for the behavioural and social sciences, launching this week, and one for ecology, evolution and environment (EEE) research, to follow later this month. Authors will be prompted to use these documents to provide important details of study design, data collection and analysis before papers are sent out for review.

In-house editors in behavioural and social sciences across the Nature journals developed the first document to address the distinct needs of research in this field — one that is remarkably broad, and includes numerous disciplines with distinct identities. Even in the same area, research protocols can vary substantially, ranging from qualitative and interpretative methods to deductive and quantitative approaches.

This presents some challenges when deciding on priorities. Most experimental approaches value sample size, for instance. But survey-based projects must also consider whether data are collected from an appropriately representative sample. Data collection methods may be relatively easy to standardize in a laboratory, but they can vary in fieldwork when scientists use a wide array of tools, or when language or literacy barriers must be overcome.

Despite these variations, our editors think it is valuable to consider all behavioural- and social-sciences research together to try to bridge the methodological divides between and within fields. We hope that describing study elements in a standardized way across the full suite of social-sciences methodologies and data types will help our multidisciplinary readers to appreciate and understand diverse research approaches.

The reporting-summary document for the behavioural and social sciences was developed on the basis of feedback from researchers with different disciplinary backgrounds and methodological expertise, including quantitative and qualitative analysis, and lab-based and field studies. It aims to capture key elements of how studies were designed, conducted and analysed — but it does not seek to enforce a specific set of standards.

For instance, determining sample size statistically using a power analysis might be best practice in experimental psychology, but it is not a necessary or sensible step in an anthropological study of a small village, which may have low sample size but effectively encompasses the entire available population. However, in both cases, authors should be able to provide a full report of how sample size was selected. Accordingly, the reporting summary is designed to be flexible.

Similar considerations motivated the design of the reporting-summary document for EEE studies. The Nature journals are among thousands of publications and organizations that have signed up to the Transparency and Openness Promotion (TOP) guidelines (B. A. Nosek *et al. Science* **348,** 1422–1425; 2015). The EEE reporting summary is being designed by in-house editors using the Tools for Transparency in Ecology and Evolution as a guide.

EEE studies include many distinctive features, and the reporting summary is being designed with those in mind. For example, there will be questions about fieldwork conditions and the treatment of wild animals, and authors of palaeontological work will be asked to describe specimen provenance, deposition and dating methods. We are currently integrating feedback from researchers into a final version.

The reporting-summary documents are a first step towards ensuring that the relevant communities pay systematic attention to reporting and transparency. These documents are not static, and the first iterations are intentionally broad. We look forward to receiving comments and thoughts. ■

KOFI ANNAN FOUNDATION

Data can help to end malnutrition across Africa

Progress in the fight against hunger is patchy. New tools must target action to those who are most vulnerable, says **Kofi Annan***.*

In 2000, the United Nations hosted the largest gathering of political leaders ever held. At that meeting, all 189 UN member states, plus leading development institutions, committed to the Millennium Development Goals, a set of eight ambitious goals for lifting more than one billion people worldwide out of extreme poverty.

The first goal — to cut extreme poverty and hunger in half by 2015 — was especially important to me, because it was crucial to achieving all the others. It was also controversial: experts thought it was impossible to achieve. But it sparked a global conversation about how to invest in agriculture, nutrition and food systems to ensure a future in which all children get the food they need to thrive, not just to survive.

Talk led to action, and action to results. Between 2000 and 2015, nearly every African country improved childhood nutrition, especially in reducing stunted growth caused by malnutrition. For example, in Burkina Faso, stunting in children younger than 5 dropped from 42% in 2006 to 27% in 2016. In Ghana, my home country, rates fell from 36% to 19% between 2003 and 2014.

These numbers are brought to life by maps produced by the University of Washington's Institute for Health Metrics and Evaluation in Seattle (see page 41). They illustrate rates of stunting, wasting and underweight in children — the best indicators for measuring child nutrition — across Africa from 2000 to 2015. These advanced statistical methods reveal progress at a level of detail that shows change almost down to the village level. A companion project has tracked childhood education, another crucial driver for improving people's lives (see page 48).

The results alone are astonishing, especially for me — an African accustomed to international headlines depicting a continent consumed by war, famine and hunger. The Africa shown in these maps tells a different story: one of measurable, steady progress on issues long thought intractable.

The maps also highlight stark disparities, particularly in conflict-affected areas. There are villages where all children are too short for their age. Across most of the Sahel, a semi-arid swath of land from the Atlantic to the Red Sea, high rates of stunting persist, with no hint of improvement.

Indeed, they are a clear reminder that national averages do not tell the full story. In Kenya, for example, rates of wasting in children under 5 were below 6% on average nationwide in 2015, yet in certain regions plagued by several years of poor rains, crop failure and disease outbreaks, estimated levels of wasting reach as high as 28%. And Chad has areas of stunting that exceed 50%, despite a national average of about 37%. In Nigeria, we see progress in the south, but stagnant and high stunting rates in the drier, conflict-ridden north.

Such fine-grained insight brings tremendous responsibility to act. It shows governments, international agencies and donors exactly where to direct resources and support. The Sustainable Development Goals — which UN member states endorsed when the Millennium Development Goals expired in 2015 — include the first targets for reducing stunting and wasting. The data indicate that no African country is currently on track to reach all the targets associated with ending hunger, achieving food security and improving nutrition.

This shows how crucial it is to invest in data. Data gaps undermine our ability to target resources, develop policies and track accountability. Without good data, we're flying blind. If you can't see it, you can't solve it.

Several nations, including Burkina Faso and Ghana, have reaped the benefits of regularly and frequently collecting data on key nutrition indicators. Importantly, they are using the data to inform decisions about policy and programmes. And countries that make nutrition a political priority are seeing results. For example, Senegal's stunting rate dropped by nearly one-third between 2011 and 2015, after the Prime Minister's office established the *Cellule de Lutte contre la Malnutrition*, a coordinating body tasked with reducing undernutrition.

WITHOUT **GOOD DATA,** WE'RE FLYING BLIND. IF YOU CAN'T **SEE IT,** YOU CAN'T **SOLVE IT.**

This progress should spark a renewed commitment to refining data collection and analysis so as to hone interventions that can reach the most vulnerable individuals: infants, children and mothers. We must apply these lessons to communities that have not fared as well.

Nutrition is one of the best drivers of development: it sparks a virtuous cycle of socio-economic improvements, such as increasing access to education and employment. With the help of institutions such as the Bill & Melinda Gates Foundation — which also supported the mapping project — I continue to advocate for better policies through my Foundation's Combatting Hunger programme. Eradicating malnutrition is crucial to delivering on the Sustainable Development Goals' promise of "leaving no one behind".

Current and former African leaders are now stepping up as part of the African Leaders for Nutrition Initiative, which was launched in January to catalyse and sustain political will. The group has committed to developing a Nutrition Accountability Score Card to track progress by country and region.

These maps are another tool in our arsenal. Alone, they won't eradicate malnutrition — but they will enable Africa's leaders to act strategically. ■

Kofi Annan *is chair of the Kofi Annan Foundation in Geneva, Switzerland; former secretary-general of the United Nations; and a Nobel Peace Prize laureate.*
e-mail: info@kofiannanfoundation.org

SEVEN DAYS *The news in brief*

EVENTS

University strike

UK academics began a strike on 22 February, over changes to their pension scheme that they say would leave them thousands of pounds worse off a year in retirement. More than 42,000 academics — members of the University and College Union — from 64 universities were called out on strike. See page 14 for more.

Gene-editing deal

US drug firm Gilead Sciences will invest up to US$3 billion in a company that specializes in gene-editing technology, it said on 22 February. Gilead subsidiary Kite will partner with Sangamo Therapeutics of Richmond, California, which has developed gene-editing techniques based on enzymes called zinc-finger nucleases. The goal is to develop cell therapies against cancer, including treatments that could be used in many people, rather than being tailor-made from an individual's own cells. Kite is headquartered in Santa Monica, California.

Turkish statement

Turkey's Science Academy has issued a rebuke to its government over what it says are "growing restraints" on freedom of expression in the nation. In a statement posted on 22 February, it says that academics who are critical of their government's policies are routinely charged with supporting terrorism, leading to dismissal from their universities or arrest — and that their claims of injustice are not properly investigated. It argues that the constitutional right to freely express opinions is the "foundation of the freedom of science", and is not legally affected by the ongoing state-of-emergency measures introduced after a failed coup attempt against the Turkey's government in July 2016.

US retreat

The US National Science Foundation plans to close all three of its overseas offices — in Beijing, Tokyo and Brussels — in the coming months. The offices, which together cost about US$2 million a year, are meant to promote collaboration between US and international researchers. The Tokyo office opened in 1960, the European office (initially in Paris) in 1984 and the Beijing one in 2006. The agency announced on 21 February that it would instead send agency experts for "short-term expeditions to selected areas to explore opportunities for collaboration".

MIKE HUTCHINGS/REUTERS

Cape Town expected to run dry in July

Officials in Cape Town, South Africa, have pushed back the date they expect the city to run out of water to 9 July, according to a 19 February announcement. Called Day Zero, the date is based on the amount of water left in reservoirs and has shifted multiple times, ranging from mid-April to early July. The city would be the first major metropolis to run dry. The temporary reprieve is the result of water restrictions that limit each person to less than 50 litres of water a day. Not all residents are abiding by this, but the weekly drop in dam levels has slowed slightly. Farmers north of Cape Town have also begun transferring some of their water to the city's reservoirs.

POLICY

Primeval forest

Increased logging in Poland's ancient Białowieża Forest breaches European Union nature laws, a leading official of the European Court of Justice (ECJ) has declared. In 2016, the Polish government tripled the logging limit in the forest — which is one of Europe's last patches of primeval forest and is protected by EU wildlife laws — on the grounds that a beetle pest needed to be controlled. Last July, the ECJ issued an interim order to stop tree-felling in strictly protected parts of the forest at Poland's border with Belarus, but logging continued. In a legal opinion published on 20 February, the ECJ's advocate-general proposes that the court should rule that the Polish government has failed to meet its obligations under the EU's habitats and birds directives. The court is expected to deliver its final judgment on the case in March.

South Africa chief

In a cabinet reshuffle on 26 February by South Africa's new president, the science and technology minister, Naledi Pandor, was moved to the post of higher-education minister. Observers hope that Pandor, who has been lauded for her contribution to research in Africa, will stabilize the country's underfunded tertiary-education system. South African universities have struggled to deal with violent protests for free education and calls for racial and curriculum change. Mmamoloko Kubayi-Ngubane, previously minister of communications, becomes science minister. She will guide the country through projects including the Square Kilometre Array, a multi-million-dollar radio telescope to be built in South Africa and Australia.

PUBLISHING

Open-access review

The United Kingdom's main public research funder will reassess its open-access policy, amid concern that a national drive to make papers free to read might not be financially sustainable. A new UK Research and Innovation body will unite nine UK research-funding agencies when it begins work in April. It will conduct an internal review of the policy this year, its chief executive, Mark Walport, said at a meeting on 20 February. Since 2013, British research councils have provided block grants totalling more than £80 million (US$112 million) to help universities pay fees to subscription publishers to make studies open-access; Walport said the policy has not met its targets.

RESEARCH

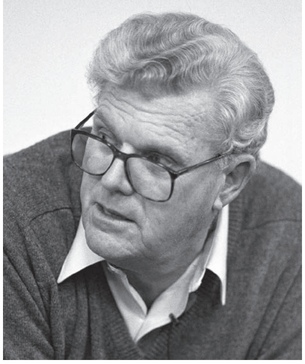
Canada survey

Just over half of Canadian government scientists say they cannot speak freely about their work, according to a survey released on 21 February by the Professional Institute of the Public Service of Canada in Ottawa. Some survey respondents said that middle managers are clinging to old restrictions that hinder the sharing of scientific findings, two years after the government of Prime Minister Justin Trudeau lifted such rules. Of about 3,000 survey respondents, 47% reported that they can speak freely to the media about their work, up from 10% in 2013. The percentage who agreed that political interference is an obstacle to using scientific evidence in decision-making dropped from 71% to 40%.

PEOPLE

Quark pioneer

Canadian-born physicist Richard Taylor, the co-discoverer of quarks, died on 22 February, aged 88. In experiments that began in the late 1960s at the Stanford Linear Accelerator Center in California, Taylor (**pictured**) and his collaborators smashed high-energy electrons into protons and found that the electrons bounced off in a surprising pattern. They later interpreted the results as confirmation of a theory proposed a few years earlier, in which particles such as protons and neutrons are composed of quarks and gluons. In 1990, Taylor shared a Nobel physics prize for the discovery.



CHUCK PAINTER/STANFORD

European chief

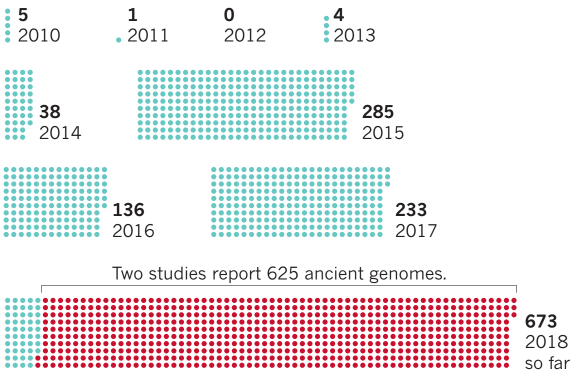
The European Commission has appointed French civil servant Jean-Eric Paquet as director-general of research and innovation. His main challenge will be to implement the ninth European Union Framework Programme (FP9), the bloc's chief science-funding mechanism, which launches in 2021. Details of the seven-year programme, whose budget is likely to exceed €100 million (US$123 million), will be announced by June. Paquet has held several commission posts over 23 years, including one in the cabinet of a previous research commissioner. He takes office on 1 April, replacing Robert-Jan Smits, who designed the current Framework programme, called Horizon 2020, and the scope of FP9.

TREND WATCH

The publication of 625 ancient-human genomes in two studies last month has nearly doubled the number of such genomes published, to about 1,300. In the past decade, the ancient-genomics field has been propelled both by improvements in sequencing technologies that can read the short pieces of DNA typically found in ancient-human remains, and by sampling of a DNA-rich inner-ear bone called the petrous — allowing researchers to do less sequencing to get the data they need.

ANCIENT-GENOMICS BOOM IN FULL SWING

The first ancient-human genome was published eight years ago. Two February studies publishing 625 genomes now push the total to more than 1,300.



SOURCE: DAVID REICH, HARVARD MEDICAL SCHOOL

FACILITIES

Arecibo shift

A consortium led by the University of Central Florida in Orlando will take over management of the Arecibo radio telescope in Puerto Rico from 1 April. It is a major shift for the historic facility, which has long been supported by the US National Science Foundation; the agency decided to divest itself of Arecibo to free up money for other astronomical facilities. The new management team includes the Metropolitan University in San Juan and Yang Enterprises of Oviedo, Florida. In announcing the decision on 22 February, the team said that it would introduce new technologies to expand the capabilities of the telescope, which not only does radio astronomy but also tracks asteroids and studies atmospheric physics.

FUNDING

Singapore budget

In Singapore's 2018 budget, the government announced that its investment in research and development (R&D) will remain at 1% of the country's gross domestic product: 4.6 billion Singapore dollars (US$3.5 billion). The budget, released on 19 February, also committed 50 million Singapore dollars to a partnership between the National Research Foundation (NRF) and state investor Temasek Holdings, with the aim of growing start-ups using intellectual property commercialized from NRF-funded research. Temasek will commit an equal sum to the scheme. The government pledged 500 million Singapore dollars for programmes to assist companies in developing and testing automated and digital technologies for the aviation and maritime industries.

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

STAFFAN WIDSTRAND/NPL

Jaguars roam from the southwestern United States to Paraguay, but their habitat has shrunk as agriculture and deforestation have expanded.

WILDLIFE TRADE

China's lust for jaguar fangs imperils big cats

Wildlife traffickers in South America seek body parts from protected species.

BY BARBARA FRASER

The jaguar was found floating in a drainage canal in Belize City, Belize, on the day after Christmas last year. Its body was mostly intact, but the head was missing its fangs. On 10 January, another cat — an ocelot that may have been mistaken for a young jaguar — turned up headless in the same channel.

The killings point to a growing illicit trade in jaguars (*Panthera onca*) that disturbs wildlife experts. The cats' fangs, skulls and hides have long been trophies for Latin American collectors who flout international bans against trading in jaguar parts. But in recent years, a trafficking route has emerged to China, where demand for jaguars could be increasing because of crackdowns on the trade in tiger parts used in Chinese traditional medicine.

Wildlife trafficking often follows Chinese construction projects in other countries, because Chinese workers can send or take objects home, says ecologist Vincent Nijman of Oxford Brookes University in Oxford, UK. "If there's a demand [in China] for large-cat parts, and that demand can be fulfilled by people living in parts of Africa, other parts of Asia or South America, then someone will step in to fill that demand," he says. "It's often Chinese-to-Chinese trade, but it's turning global."

That seems to be the case in Bolivia, where 8 packages containing a total of 186 jaguar fangs were confiscated between ▶

August 2014 and February 2015 before they could reach China. Seven had been sent by Chinese citizens living in Bolivia. Eight more were reportedly intercepted in 2016, and a package of 120 fangs was seized in China, says Angela Núñez, a Bolivian biologist who is researching the trade.

Those packages could represent the deaths of more than 100 jaguars, although it's impossible to be sure, Núñez says. In northern Bolivia, where several Chinese companies are working, radio advertisements and flyers have offered US$120 to $150 per fang — more than a month's income for many local people. Two Chinese men have been arrested for trading in jaguar parts. One, detained in 2014, received a three-year suspended sentence. The other, arrested in 2016, is awaiting sentencing but failed to appear for two recent court hearings.

Worldwide, very few wildlife-trafficking cases lead to criminal sentences, Nijman says. "The deterrent is when somebody ends up in jail," he says — but that rarely happens.

Fangs and skulls seized in Bolivia, as well as 38 fangs confiscated in Lima, Peru, in 2015, could have come from jaguars that were killed recently, or years ago. Because the cats have large territories, Núñez says that genetic studies could determine whether poached animals came from populations in Bolivia.

That also interests Brazilian biologist Thais Morcatty, who is doing her PhD research with Nijman. There is a domestic market in Brazil for jaguar skins as home decoration, but parts of the animals have also been shipped abroad from Rio de Janeiro and São Paulo, she says.

More than a century ago, jaguars roamed forests, savannahs and scrub land from the southwestern United States to Paraguay. Deforestation and other disturbances caused by people — especially the expansion of agriculture — have cut the cats' habitat in half, says wildlife ecologist John Polisar, who coordinates the jaguar programme at the Wildlife Conservation Society in New York City.

"It's often Chinese-to-Chinese trade, but it's turning global."

That has depleted the jaguars' prey, and in some areas has forced the cats into contact with people and livestock, says Polisar, who works in Central and South America. Estimates of the remaining jaguar population range from about 60,000 to nearly three times that number.

A farmer who loses a cow or calf to a predator might kill a jaguar in retaliation, even though that animal might be innocent. After habitat loss, such killings are the second-biggest threat to jaguars, says Esteban Payán, director of the northern South America jaguar programme at Panthera, a global wild-cat conservation organization. The retaliatory killings also provide a sporadic supply of animal parts to the wildlife trade, but sparse data make it difficult to know whether the incidents are isolated cases.

Measures designed to help people coexist with jaguars could reduce such killings, Payán says. In some cases, electric fences have discouraged jaguars from crossing from forests into pastures, and solar panels that power the fences can also run some light bulbs or a small refrigerator for the farmer's family (H. Quigley *et al. PARKS* **21.1,** 63–72; 2015). That can revolutionize life for them, he says.

Other tactics that have shown promise include putting bells on cows and installing flashing lights around pastures to help keep predators at bay. Introducing guard animals to a herd, such as burros (a type of donkey), can also discourage predators, he says.

Governments could help by providing incentives, says biologist Ricardo Moreno, director of the non-profit group Yaguará Panama. Now, a farmer who buys a cow on credit must repay even if he loses an animal, says Moreno, who mixes scientific studies and work with communities and policymakers to protect jaguars. But making loans contingent on better livestock management would benefit farmers, lenders and jaguars, he says.

Meanwhile, researchers and some government officials in Latin America are watching the wildlife trade warily. Belize's environment ministry is offering a US$5,000 reward for information about the jaguars killed there, and Polisar's group is collecting data from around the region.

Although the links to international trafficking in Bolivia are clear, Payán worries this is "just the tip of the iceberg" of a broader trading network, because there are anecdotal reports of trafficking in other countries. Conservation groups are no match for "the violence, the money and the scale" of organized poaching rings, he says. "The potential threat is huge." ■

UNIVERSITY WALKOUT

UK universities cope with disruption from huge strike

Pension changes spur more than 40,000 academics to walk out on research and lectures.

BY ELIZABETH GIBNEY

Britain's leading research universities are coping with the disruption wrought by a nationwide strike, as academics protest against changes to their pensions. The walkout, which began on 22 February, is one of the largest by university staff in the country's recent history and is disrupting scientific experiments, conferences and lectures.

More than 42,000 academics — members of the University and College Union (UCU) — were called out on strike from 64 institutions across the United Kingdom. About 25,500 of those members are research staff, and the dearth of lecturers is predicted to affect more than 1 million students. Fourteen days of strikes are scheduled over four weeks.

Academics are walking out over planned changes to the Universities Superannuation Scheme, the main pension fund for 190,000 faculty members and staff at many of Britain's older, research-intensive universities. A 2017 valuation found that the fund had a growing deficit of £12.6 billion (US$17.6 billion) — one of the largest of any private UK pension scheme. Universities UK, which represents the academic employers, says that the fund will be difficult to sustain without reform. It proposed changes — pushed through in January — that would see pension income go from having a guaranteed element to being entirely dependent on investment return. According to financial models commissioned by Universities UK, pension recipients would lose £2,000–5,000 of income a year, depending on salary. The UCU puts the loss at as much as £10,000 a year, and says that the proposals are based on an overly pessimistic view of the fund's deficit. That position is backed by a growing number of UK institute heads, who have broken ranks on the issue.

As *Nature* went to press, the two sides were expected to begin fresh talks on 27 February. Universities UK told pension-scheme members in a letter that it would be open to reintroducing

XED.
www.ucu.org.uk/strikeforuss
OUR PENSION AXED. Support our strike to defend USS
March for education Back pension & pay strikes
London, Eat & South Eas

Academics on the picket line at University College London.

LIZZY BROWN/NATURE

guaranteed benefits if economic conditions improve. But the group said that the upcoming talks would not reopen the January decision to move forward with the changes. The UCU has agreed to discussions, but said that strikes this week would continue because that decision was the "very reason" for the walkouts.

Academics already face pay rises below inflation, an insecure career path and increasing workloads, Sally Hunt, general-secretary of the London-based UCU, said at a press conference. "It has always been understood that part of the package that they could look forward to was a reasonable pension," she says. The action is "unprecedented", says Hunt. "I can't within my time in UCU remember anything as serious."

Striking staff are not doing research or attending tutorials, lectures or external commitments. Cancelled lectures will not be rescheduled, says the union. The first strike period lasts for five days, and four- and five-day phases are scheduled to follow. The UCU estimates that, across the 14 days, 575,000 teaching hours will be lost. More than 110,000 students have signed petitions calling for financial compensation for missed teaching.

SCIENCE TO THE SIDE

Ian Gent, a computer scientist at the University of St Andrews, says that the strike could stop his team from bidding to host a doctoral-training centre in artificial intelligence, worth around £5 million. UK funding agencies announced the opportunity two weeks ago, with a short deadline. "It would be no surprise if we could not write the bid on time," says Gent. But the future well-being of staff is in jeopardy, he says, and academics must stand up against that.

The action will no doubt affect research, says Aimee Grant, who studies public health at Cardiff University and is on a research-only contract. She will lose 14 days of work on her current project on breastfeeding in public spaces. Researchers with contracts such as hers, which are based on completing a defined project, will have to catch up out of hours, she says. But Grant urges other research-only staff like her to strike.

"We hope that employees recognise that changes are necessary to put the scheme on a secure footing, and that the proposed strike action will only serve to unfairly disrupt students' education," Universities UK said in a statement. It added that it was doing everything it could to minimize disruption. Cancelled academic conferences include a seminar on archaeology and genetics at University College London and one on the Rohingya refugee crisis at SOAS University of London.

> *"I can't within my time in UCU remember anything as serious."*

The strike comes after formal negotiations between the two sides ended in January. A narrow majority of the joint negotiating committee sided with Universities UK and approved the group's proposals to address the deficit; none of the five UCU committee members voted in favour of them. The changes are subject to a routine consultation period, during which Universities UK will discuss the plans with employers and affected employees, but they are not obliged to amend the proposals in response.

Earlier in January, the UCU had balloted its members on their willingness to strike should talks end unsuccessfully. Fifty-eight per cent voted, and 88% of them backed strike action. Universities UK estimates that those voting in favour of the strike account for 16% of academic staff at UCU-represented institutions. ■

POLICY

Florida bills to impact schools

Residents could influence classroom materials.

BY GIORGIA GUGLIELMI

Policymakers in the United States are pushing to give the public more power to influence what educators teach students. Florida's legislature has started considering two related bills that, if signed into law, would let residents recommend which instructional materials teachers in their school district use in class.

The bills build on a law enacted in June 2017, which enables any Florida resident to challenge the textbooks and other educational tools used in their district as being biased or inaccurate.

But the bills, approved in mid-February by the education committees in the state's Senate and House of Representatives, go a step further, allowing the public to review educational materials and to suggest alternatives. The final decision on whether to follow the recommendations still rests with the school boards.

Attempts to influence what students learn typically tackle the issue head-on, by trying to change state education standards. A bill introduced in Iowa on 12 February would remove guidelines in the state's science education standards to teach evolution and the effects of human activity on climate.

Florida's bills could alter classroom content in a less direct way. Allowing taxpayers to have a say in what goes on in public schools seems innocuous, says Brandon Haught, an environmental science teacher in Orange City, Florida. But the bills, together with last year's law, expose schools to activists who oppose the teaching of topics such as evolution and global warming, he says.

What's more, the bills and the law use language that makes it easier for individuals to target such topics, says Glenn Branch, deputy director of the National Center for Science Education in Oakland, California. The documents state that educational materials should be "balanced" and "noninflammatory", but they don't specify who decides whether something is inflammatory, he says.

State Representative Byron Donalds (Republican), who sponsored last year's law and this year's bill in the House, says it's important that school boards consider different viewpoints. "You can debate on things and draw your own conclusions," Donalds says.

The bills must still be voted on by the full House and Senate, but Branch says that they have a good chance of becoming law in Florida. ■

INDIA

Harassment case opens dialogue

Researchers in India say high-profile sacking should encourage women to report sexual harassment, and lead academic institutions to deal with cases fairly and fearlessly.

BY T. V. PADMA

Female scientists in India hope that more academics will be able to report sexual harassment without fear of jeopardizing their career or reputation, after a prominent biologist was sacked for allegedly harassing a staff member.

Immunologist Kanury Rao was dismissed in January from his post as national head of the Translational Health Science and Technology Institute (THSTI) at Faridabad near New Delhi. The move followed an investigation by the institute's internal complaints committee (ICC) in mid-2017, which found that Rao "used unwelcome sexually determined behaviour" towards a junior female staff member at the institute, harassing her and interfering with her work in incidents between late 2014 and 2017. The ICC said that Rao's behaviour violated several Indian Central Civil Services Rules, which cover public research institutes such as the THSTI.

Rao denies the allegations, and told *Nature* that he filed an appeal last week with the institute's appellate authority to dispute the sexual-harassment allegation and the termination of his employment. "The allegations of sexual harassment against me were entirely fabricated and simply represent a case where professional disgruntlement evolved into a larger conspiracy to malign and defame me, in order to eventually effect my removal," Rao says. "Unfortunately, the inquiry committee also solely went by empirical impressions, relying for its findings only on weak circumstantial criteria rather than any real evidence." The THSTI did not respond to a request for comment about Rao's assertions.

The complainant, who wishes to remain anonymous, denies Rao's allegations. After a thorough investigation, the ICC determined that the senior scientist's behaviour was unacceptable, she says.

"What professional disgruntlement could be so big that a young girl would challenge a senior and highly influential scientist?" she says. "It is not easy for a young woman to come forward and open up about these kind of issues."

Some scientists who spoke to *Nature* say there have been few cases in which a scientist in India has been sacked for sexual harassment. "I am not aware of a previous case of any Indian scientist, let alone a high-profile one, being dismissed on sexual-harassment charges," says Rahul Siddharthan, a computational biologist at the Institute of Mathematical Sciences in Chennai.

Beyond Rao's case, says immunologist Vineeta Bal of the Indian Institute of Science Education and Research in Pune, sexual harassment of women in science in India is not uncommon. "What is uncommon is somebody coming forward to report it and seek justice," she says. Several Indian scientists have told *Nature* that female researchers who have experienced sexual harassment still seem to be reluctant to speak out. "Most women fear putting their jobs, career and reputation in jeopardy by going public," says Bal.

Evidence suggests that sexual harassment is pervasive in India. A 2017 survey commissioned by the Indian National Bar Association found that of 6,047 respondents, 47% had experienced unwelcome sexual comments, jokes or gestures at work (see 'Workplace harassment').

Many workplaces, including scientific institutions, have mechanisms for investigating allegations of sexual assault or harassment, such as internal complaints committees. But Siddharthan says that, in practice, some of these committees have not been very proactive. The Bar Association's report included an in-depth survey of 45 people who had been sexually harassed; only one-third felt that complaints committees had dealt with their cases fairly. "One can hope that this [Rao] case will change things, and ICCs are encouraged to deal with such cases fairly and fearlessly," says Siddharthan.

Biologist Vidita Vaidya at the Tata Institute of Fundamental Research (TIFR) in Mumbai says another problem that makes some women reluctant to report workplace harassment is the expectation that they will have to endure a difficult and intrusive inquiry. "There is a real fear for many that rather than their work and scientific merit defining their trajectory and career, this complaint will become a focal point for all discussions regarding them," says Vaidya, a former member of a TIFR support group called the women's cell.

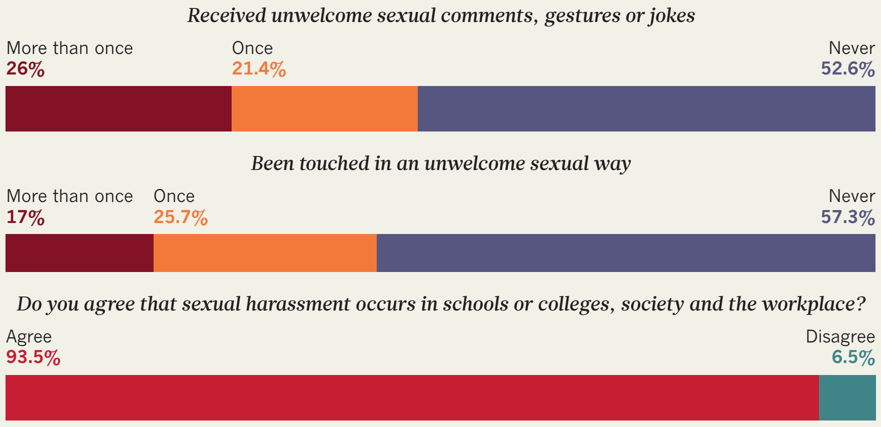
Vaidya says that some women prefer to keep their harassment quiet because they feel they can cope by removing themselves from the situation. "Cases where complaints have been made are simply the tip of the iceberg," she says.

Vaidya worries that some women might be walking away from science to avoid workplace harassment. More people might be inclined to report sexual harassment if research institutions explained what happens when a complaint is made, she says. "Clearly indicating that their confidentiality will be protected will go a long way [to improve] their faith in the system," she says.

Other scientists sense that change is already on the way. Bal says that more women in academia might be encouraged to come forward with complaints and expect justice as a result of the publicity generated by high-profile cases, such as the many women in the entertainment industry who say they have been assaulted or harassed. "When institutions send a clear message that complaints will be heard and acted upon, this should give courage to people who face harassment to approach the [women's] cell and speak out," says biologist Sandhya Koushika, chair of the TIFR women's cell. ■

WORKPLACE HARASSMENT

A 2017 survey commissioned by the Indian National Bar Association received responses from 6,047 men and women about sexual harassment at work. It found that many had experienced some form of harassment.



SOURCE: INDIAN NATL BAR ASSOC. (2017); HTTP://GO.NATURE.COM/2F14I3I

In the event of an emergency, contact / Uma kuphuthuma, ungashayela:
Pharmacist of Record: Aruna Bhoola: 082 777 0903 (REG. NO. 15262)
Pharmacy Manager: Jayajothi Moodley: 082 659 3766 (REG.NO. 11622)
SAMRC: HPRU – VERULAM CRS
STOP
VACCINE PREPARATION IN PROGRESS: PLEASE DO NOT DISTURB
DISPENSING AREA

Researchers around the world are conducting 12 late-stage trials of HIV vaccines.

JACKIE CLAUSEN/THE TIMES/GALLO/GETTY

PUBLIC HEALTH

HIV–vaccine strategy sought

Therapies to prevent infection advance in a crowded field.

BY AMY MAXMEN

Several vaccines and drugs for preventing the spread of HIV are showing signs of success in clinical trials, three decades after scientists began the search. But some researchers fear that progress will stall without a coordinated strategy to ensure that the most promising therapies to prevent infection win support from policymakers and reach the people who need them.

A meeting convened by the World Health Organization (WHO) in Geneva, Switzerland, from 28 February to 1 March aims to address a lack of long-term thinking about the factors — such as cost and ease of use — that can determine whether a vaccine or other preventive therapy succeeds in reducing disease. Some HIV researchers argue that they should study these issues now, while clinical trials of potential vaccines and drugs are ongoing, to avoid delays in delivering effective therapies to people at risk of infection. Many hope that the WHO meeting will trigger broader discussions about how to support such research given limited resources, and how to prioritize therapies in development.

Waiting to conduct these kinds of studies until trials are finished prolongs the time for a preventive therapy to reach people. In the meantime, the epidemic worsens. Worldwide, about 1.8 million people contracted HIV in 2016. “You need to have a good idea about where you want to end up and all of the steps you need to make to get there,” says Mark Feinberg, president of the International AIDS Vaccine Initiative in New York City.

But it is not clear who would make decisions about which projects to prioritize, or when the choices would be made.

Some 25,000 people around the world are participating in clinical trials of treatments to prevent HIV infection. Twelve late-stage trials worldwide are testing experimental vaccines; these include a 2,600-person study in southern Africa of a vaccine designed to block multiple strains of the virus. Others are assessing the potential of proteins called broadly neutralizing antibodies, which might stop HIV from infecting immune cells. And a pair of phase III trials has enrolled 7,700 people to test whether injections of the drug cabotegravir can prevent HIV infection for two months at a time.

DELIVERY CONCERNS

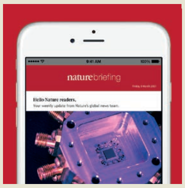
At the meeting, researchers, policymakers and HIV activists will discuss stumbling blocks that have limited the use of potent vaccines and treatments against other diseases, such as high costs and cumbersome delivery requirements. Because no therapy has approached 100% protection against HIV, regulators face tough decisions when considering the cost and effort of delivering treatment to people at risk. In 2009, for example, a phase III study of the most promising vaccine identified so far found that it reduced a person’s risk of contracting HIV by only one-third (S. Rerks-Ngarm *et al. N. Engl. J. Med.* **361,** 2209–2220; 2009). Health authorities did not recommend it for widespread use.

A modified version of that vaccine is now being tested in 5,400 people in South Africa, and researchers hope that it will reduce a person’s chance of contracting HIV by at least 50%. But even if the trial succeeds, the expense and difficulty of administering the vaccine, which must be given as six injections over 18 months, could make it a hard sell to policymakers and funders. Health-care workers around the world struggle to persuade healthy people to get one-time shots that are highly effective against other deadly diseases.

Similar concerns surround the antibodies in development, because they are given as intravenous infusions, and it is unclear how long treatment must continue to prevent HIV. The antibodies are also relatively expensive to make. Eventually, scientists must be prepared to choose which projects to stall, and which to supplement with studies aimed at developing cheaper, easier ways of administering a given therapy, says Mitchell Warren, executive director of AVAC, an HIV-prevention advocacy ▶

MORE ONLINE

NATURE BRIEFING



Save time — get the daily *Nature* newsletter free in your inbox go.nature.com/2bhx2ks

MORE NEWS

- ● Doubts raised over Australia’s plan to wipe out carp go.nature.com/2svpywi
- ● Amateur astronomer catches birth of a supernova go.nature.com/2bvbvct
- ● Neanderthal artists made oldest-known cave paintings go.nature.com/2eudrai

NATURE PODCAST

Childhood cancers; a fingerprint from the early Universe; and brain waves cause a splash nature.com/nature/podcast organization in New York City.

Money is limited, as is the pool of people available for clinical trials, which become larger and more complex as a vaccine or antibody treatment progresses towards the market. “We will need prioritization,” Warren says. “That view needs to be driven by science and financial realities, and the decision process needs to be clear and transparent.”

Another issue facing researchers is how to improve the likelihood that people at risk of HIV infection will take preventive treatments. Success is not guaranteed: Truvada, a daily pill for preventing HIV infection, has not reduced the number of new HIV cases globally since regulators approved it six years ago. In eastern and southern Africa, for instance, young women rarely take the drug, even though they account for 26% of the region’s new infections. Tian Johnson, founder of the African Alliance for HIV Prevention in Johannesburg, South Africa, says that researchers did not adequately consider how poverty, pregnancy, discrimination and abuse might affect whether young women at risk are likely to seek out Truvada. “If you disregard the complexity of a woman’s daily life and reality, you put at risk the millions of dollars you invest in developing a product,” Johnson says.

Despite the challenges ahead, the fact that these discussions are happening is an important step forward, says Feinberg. “You can’t keep your head in the sand,” he says. “You need to work ahead and think of ways that we as a research-development community can solve these problems — and they are solvable.” ■

PUBLISHING

Duplicated images could soon be identified by an automated test

Team says technique finds reused images even if they have been rotated and resized.

BY DECLAN BUTLER

Researchers have developed an automated technique that they say can quickly detect duplicate images among hundreds of thousands of papers. If it proves successful, the software could make it easy for editors to screen images before publication — something that currently requires great effort and is done by only a few publications.

Daniel Acuna, a machine-learning researcher at Syracuse University in New York, and his two colleagues described their algorithm on 22 February (D. E. Acuna *et al.* Preprint at bioRxiv http://dx.doi.org/10.1101/269415; 2018).

Acuna says he isn’t making the full algorithm public, because that could trigger false allegations. Instead, his team plans to license it to journals and research-integrity offices. Lauran Qualkenbush, director of the Office for Research Integrity at Northwestern University in Chicago, Illinois, and vice-president of the US Association of Research Integrity Officers, says she has discussed the approach with Acuna. “It would be extremely helpful for a research-integrity office,” she says.

In early 2015, Acuna’s team used the algorithm to extract more than 2.6 million images from the 760,000 articles then in the open-access subset of the PubMed database of biomedical literature. These included micrographs of cells and tissues, and gel blots. The algorithm then zoomed in on the most feature-rich areas — where colour and greyscales vary most — to extract a characteristic digital ‘fingerprint’ of each image.

The researchers only compared images across papers from the same first and corresponding authors, to avoid the computational load of comparing every image against every other one. But the system could pick up potential duplicates even if they had been rotated, resized or had their contrast or colours changed. The trio then manually examined a sample of around 3,750 of the flagged images to judge whether the duplicates were suspicious or potentially fraudulent. On the basis of their results, they predict that 1.5% of the papers in the database would contain suspicious images, and that 0.6% of the papers would contain fraudulent images.

“It would be extremely helpful for a research-integrity office.”

The researchers haven’t been able to benchmark the accuracy of their algorithm, says Hany Farid, a computer scientist at Dartmouth College in Hanover, New Hampshire — because there isn’t a database of known duplicate or non-duplicate scientific images against which they could test the tool.

At present, many journals check some images, but relatively few have automated processes. For instance, *Nature* runs random spot checks on images in submitted manuscripts. (*Nature*’s news team is editorially independent of its journal team.)

To detect image reuse across the literature, publishers would need to create a shared database of all published images against which articles submitted for publication could be compared, says IJsbrand Jan Aalbersberg, head of research integrity at the Dutch publishing giant Elsevier.

There are currently no plans for a publisher-wide system for image checking, but that is partly because the technologies are not yet mature, says Ed Pentz, executive director of Crossref, a non-profit collaboration of 10,000 publishers. Crossref runs a service that enables publishers to routinely screen submitted manuscripts for plagiarism.

Elsevier says it would support such an initiative for images. Two years ago, the company set up a 3-year, €1-million (US$1.2-million) partnership with Humboldt University in Berlin to study article mining and to identify research misconduct. On 25 January, the project announced that it intends to create a database of images from retracted publications. Such a data set would provide a bank of test images for researchers developing automated screening of images in publications. ■

CORRECTION

In saying that everyday atomic hearts have equal protons and neutrons, the News story ‘Physicists plan first antimatter road trip’ (*Nature* **554,** 412–413; 2018) didn’t take account of the fact that some elements, such as hydrogen and lithium, have uneven numbers of protons in their most abundant form.

The News Feature ‘The entangled web’ (*Nature* **554,** 289–292; 2018) misstated the leadership of the Dutch demonstration quantum network. The project is co-led by Ronald Hanson and Stephanie Wehner of Delft University of Technology in the Netherlands and Erwin van Zwet at the Dutch research organization TNO in The Hague.

ILLUSTRATION BY PAWEŁ JOŃCA

WAVE THERAPY

How flashing lights, pink noise or other non-invasive approaches to taming brainwaves might one day turn into treatments for neurodegenerative disease.

BY HELEN THOMSON

In March 2015, Li-Huei Tsai set up a tiny disco for some of the mice in her laboratory. For an hour each day, she placed them in a box lit only by a flickering strobe. The mice — which had been engineered to produce plaques of the peptide amyloid-β in the brain, a hallmark of Alzheimer's disease — crawled about curiously. When Tsai later dissected them, those that had been to the mini dance parties had significantly lower levels of plaque than mice that had spent the same time in the dark[1].

Tsai, a neuroscientist at Massachusetts Institute of Technology (MIT) in Cambridge, says she checked the result; then checked it again. "For the longest time, I didn't believe it," she says. Her team had managed to clear amyloid from part of the brain with a flickering light. The strobe was tuned to 40 hertz and was designed to manipulate the rodents' brainwaves, triggering a host of biological effects that eliminated the plaque-forming proteins. Although promising findings in mouse models of Alzheimer's disease have been notoriously difficult to replicate in humans, the experiment offered some tantalizing possibilities. "The result was so mind-boggling and so robust, it took a while for the idea to sink in, but we knew we needed to work out a way of trying out the same thing in humans," Tsai says.

Scientists identified the waves of electrical activity that constantly ripple through the brain almost 100 years ago, but they have struggled to assign these oscillations a definitive role in behaviour or brain function. Studies have strongly linked brainwaves to memory consolidation during sleep, and implicated them in processing sensory inputs and even coordinating consciousness. Yet not everyone is convinced that brainwaves are all that meaningful. “Right now we really don’t know what they do,” says Michael Shadlen, a neuroscientist at Columbia University in New York City.

Now, a growing body of evidence, including Tsai’s findings, hint at a meaningful connection to neurological disorders such as Alzheimer’s and Parkinson’s diseases. The work offers the possibility of forestalling or even reversing the damage caused by such conditions without using a drug. More than two dozen clinical trials are aiming to modulate brainwaves in some way — some with flickering lights or rhythmic sounds, but most through the direct application of electrical currents to the brain or scalp. They aim to treat everything from insomnia to schizophrenia and premenstrual dysphoric disorder.

Tsai’s study was the first glimpse of a cellular response to brainwave manipulation. “Her results were a really big surprise,” says Walter Koroshetz, director of the US National Institute of Neurological Disorders and Stroke in Bethesda, Maryland. “It’s a novel observation that would be really interesting to pursue.”

A POWERFUL WAVE

Brainwaves were first noticed by German psychiatrist Hans Berger. In 1929, he published a paper[2] describing the repeating waves of current he observed when he placed electrodes on people’s scalps. It was the world’s first electroencephalogram (EEG) recording — but nobody took much notice. Berger was a controversial figure who had spent much of his career trying to identify the physiological basis of psychic phenomena. It was only after his colleagues began to confirm the results several years later that Berger’s invention was recognized as a window into brain activity.

Neurons communicate using electrical impulses created by the flow of ions into and out of each cell. Although a single firing neuron cannot be picked up through the electrodes of an EEG, when a group of neurons fires again and again in synchrony, it shows up as oscillating electrical ripples that sweep through the brain.

Those of the highest frequency are gamma waves, which range from 25 to 140 hertz. People often show a lot of this kind of activity when they are at peak concentration. At the other end of the scale are delta waves, which have the lowest frequency — around 0.5 to 4 hertz. These tend to occur in deep sleep (see ‘Rhythms of the mind’).

At any point in time, one type of brainwave tends to dominate, although other bands are always present to some extent. Scientists have long wondered what purpose, if any, this hum of activity serves, and some clues have emerged over the past three decades. For instance, in 1994, discoveries in mice indicated that the distinct patterns of oscillatory activity during sleep mirrored those during a previous learning exercise[3]. Scientists suggested that these waves could be helping to solidify memories.

Brainwaves also seem to influence conscious perception. Randolph Helfrich at the University of California, Berkeley, and his colleagues devised a way to enhance or reduce gamma oscillations of around 40 hertz using a non-invasive technique called transcranial alternating current stimulation (tACS). By tweaking these oscillations, they were able to influence whether a person perceived a video of moving dots as travelling vertically or horizontally[4].

The oscillations also provide a potential mechanism for how the brain creates a coherent experience from the chaotic symphony of stimuli hitting the senses at any one time, a puzzle known as the ‘binding problem’. By synchronizing the firing rates of neurons responding to the same event, brainwaves might ensure that the all of the relevant information relating to one object arrives at the correct area of the brain at exactly the right time. Coordinating these signals is the key to perception, says Robert Knight, a cognitive neuroscientist at the University of California, Berkeley, “You can’t just pray that they will self-organize.”

> *“I used to tell people — if you’re going to get Alzheimer’s, first become a mouse.”*

HEALTHY OSCILLATIONS

But these oscillations can become disrupted in certain disorders. In Parkinson’s disease, for example, the brain generally starts to show an increase in beta waves in the motor regions as body movement becomes impaired. In a healthy brain, beta waves are suppressed just before a body movement. But in Parkinson’s disease, neurons seem to get stuck in a synchronized pattern of activity. This leads to rigidity and movement difficulties. Peter Brown, who studies Parkinson’s disease at the University of Oxford, UK, says that current treatments for the symptoms of the disease — deep-brain stimulation and the drug levodopa — might work by reducing beta waves.

People with Alzheimer’s disease show a reduction in gamma oscillations[5]. So Tsai and others wondered whether gamma-wave activity could be restored, and whether this would have any effect on the disease.

They started by using optogenetics, in which brain cells are engineered to respond directly to a flash of light. In 2009, Tsai’s team, in collaboration with Christopher Moore, also at MIT at the time, demonstrated for the first time that it is possible to use the technique to drive gamma oscillations in a specific part of the mouse brain[6].

Tsai and her colleagues subsequently found that tinkering with the oscillations sets in motion a host of biological events. It initiates changes in gene expression that cause microglia — immune cells in the brain — to change shape[1]. The cells essentially go into scavenger mode, enabling them to better dispose of harmful clutter in the brain, such as amyloid-β. Koroshetz says that the link to neuroimmunity is new and striking. “The role of immune cells like microglia in the brain is incredibly important and poorly understood, and is one of the hottest areas for research now,” he says.

If the technique was to have any therapeutic relevance, however, Tsai and her colleagues had to find a less-invasive way of manipulating brainwaves. Flashing lights at specific frequencies has been shown to influence oscillations in some parts of the brain, so the researchers turned to strobe lights. They started by exposing young mice with a propensity for amyloid build-up to flickering LED lights for one hour. This created a drop in free-floating amyloid, but it was temporary, lasting less than 24 hours, and restricted to the visual cortex.

To achieve a longer-lasting effect on animals with amyloid plaques, they repeated the experiment for an hour a day over the course of a week, this time using older mice in which plaques had begun to form. Twenty-four hours after the end of the experiment, these animals showed a 67% reduction in plaque in the visual cortex compared with controls. The team also found that the technique reduced tau protein, another hallmark of Alzheimer’s disease.

Alzheimer’s plaques tend to have their earliest negative impacts on the hippocampus, however, not the visual cortex. To elicit oscillations where they are needed, Tsai and her colleagues are investigating other techniques. Playing rodents a 40-hertz noise, for example, seems to cause a decrease in amyloid in the hippocampus — perhaps because the hippocampus sits closer to the auditory cortex than to the visual cortex.

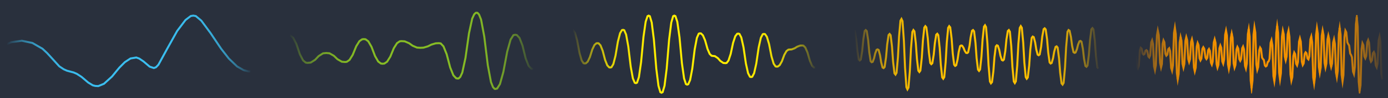
Tsai and her colleague Ed Boyden, a neuroscientist at MIT, have now formed a company, Cognito Therapeutics in Cambridge, to test similar treatments in humans. Last year, they started a safety trial, which involves testing a flickering light device, worn like a pair of glasses, on 12 people with Alzheimer’s.

Caveats abound. The mouse model of Alzheimer’s disease is not a perfect reflection of the disorder, and many therapies that have shown promise in rodents have failed in humans. “I used to tell people — if you’re going to get Alzheimer’s, first become a mouse,” says Thomas Insel, a neuroscientist and psychiatrist who led the US National Institute of Mental Health in Bethesda, Maryland, from 2002 until 2015.

Others are also looking to test how manipulating brainwaves might help people with Alzheimer’s disease. “We thought Tsai’s study

RHYTHMS OF THE MIND

Brain oscillations are characterized by their frequency, amplitude and source. Although many types of wave may be coursing through the brain at any given time, certain types dominate during particular behaviours, suggesting some mechanistic links.



Delta
0.5–4 Hz
The slowest brainwaves are associated with deep, often dreamless sleep.

Theta
4–8 Hz
In the cortex of the brain, these are seen in young children and adults in a drowsy, meditative or pathological state.

Alpha
8–13 Hz
Arising in the occipital lobe, alpha waves are associated with wakeful rest with eyes closed.

Beta
13–32 Hz
These are associated with normal wakeful consciousness and concentration, and are suppressed during movement.

Gamma
25–140 Hz
Linked to normal visual consciousness and rapid-eye-movement sleep, these might help to decipher multiple sensory signals.

SOURCE: H. MARZBANI, H. R. MARATEB & M. MANSOURIAN *BASIC CLIN. NEUROSCI.* **7,** 143–158 (2016)

was outstanding," says Emiliano Santarnecchi at Harvard Medical School in Boston, Massachusetts. His team had already been using tACS to stimulate the brain, and he wondered whether it might elicit stronger effects than a flashing strobe. "This kind of stimulation can target areas of the brain more specifically than sensory stimulation can — after seeing Tsai's results, it was a no-brainer that we should try it in Alzheimer's patients."

His team has begun an early clinical trial in which ten people with Alzheimer's disease receive tACS for one hour daily for two weeks. A second trial, in collaboration with Boyden and Tsai, will look for signals of activated microglia and levels of tau protein. Results are expected from both trials by the end of the year.

Knight says that Tsai's animal studies clearly show that oscillations have an effect on cellular metabolism — but whether the same effect will be seen in humans is another matter. "In the end, it's data that will win out," he says.

The studies may reveal risks, too. Gamma oscillations are the type most likely to induce seizures in people with photosensitive epilepsy, says Dora Hermes, a neuroscientist at Stanford University in California. She recalls a famous episode of a Japanese cartoon that featured flickering red and blue lights, which induced seizures in some viewers. "So many people watched that episode that there were almost 700 extra visits to the emergency department that day."

A BRAIN BOOST

Nevertheless, there is clearly a growing excitement around treating neurological diseases using neuromodulation, rather than pharmaceuticals. "There's pretty good evidence that by changing neural-circuit activity we can get improvements in Parkinson's, chronic pain, obsessive–compulsive disorder and depression," says Insel. This is important, he says, because so far, pharmaceutical treatments for neurological disease have suffered from a lack of specificity. Koroshetz adds that funding institutes are eager for treatments that are innovative, non-invasive and quickly translatable to people.

Since publishing their mouse paper, Boyden says, he has had a deluge of requests from researchers wanting to use the same technique to treat other conditions. But there are a lot of details to work out. "We need to figure out what is the most effective, non-invasive way of manipulating oscillations in different parts of the brain," he says. "Perhaps it is using light, but maybe it's a smart pillow or a headband that could target these oscillations using electricity or sound." One of the simplest methods that scientists have found is neurofeedback, which has shown some success in treating a range of conditions, including anxiety, depression and attention-deficit hyperactivity disorder. People who use this technique are taught to control their brainwaves by measuring them with an EEG and getting feedback in the form of visual or audio cues.

Phyllis Zee, a neurologist at Northwestern University in Chicago, Illinois, and her colleagues delivered pulses of 'pink noise' — audio frequencies that together sound a bit like a waterfall — to healthy older adults while they slept. They were particularly interested in eliciting the delta oscillations that characterize deep sleep. This aspect of sleep decreases with age, and is associated with a decreased ability to consolidate memories.

So far, her team has found that stimulation increased the amplitude of the slow waves, and was associated with a 25–30% improvement in recall of word pairs learnt the night before, compared with a fake treatment[7]. Her team is midway through a clinical trial to see whether longer-term acoustic stimulation might help people with mild cognitive impairment.

Although relatively safe, these kinds of technologies do have limitations. Neurofeedback is easy to learn, for instance, but it can take time to have an effect, and the results are often short-lived. In experiments that use magnetic or acoustic stimulation, it is difficult to know precisely what area of the brain is being affected. "The field of external brain stimulation is a little weak at the moment," says Knight. Many approaches, he says, are open loop, meaning that they don't track the effect of the modulation using an EEG. Closed loop, he says, would be more practical. Some experiments, such as Zee's and those involving neurofeedback, already do this. "I think the field is turning a corner," Knight says. "It's attracting some serious research."

In addition to potentially leading to treatments, these studies could break open the field of neural oscillations in general, helping to link them more firmly to behaviour and how the brain works as a whole.

Shadlen says he is open to the idea that oscillations play a part in human behaviour and consciousness. But for now, he remains unconvinced that they are directly responsible for these phenomena — referring to the many roles people ascribe to them as "magical incantations". He says he fully accepts that these brain rhythms are signatures of important brain processes, "but to posit the idea that synchronous spikes of activity are meaningful, that by suddenly wiggling inputs at a specific frequency, it suddenly elevates activity onto our conscious awareness? That requires more explanation."

Whatever their role, Tsai mostly wants to discipline brainwaves and harness them against disease. Cognito Therapeutics has just received approval for a second, larger trial, which will look at whether the therapy has any effect on Alzheimer's disease symptoms. Meanwhile, Tsai's team is focusing on understanding more about the downstream biological effects and how to better target the hippocampus with non-invasive technologies.

For Tsai, the work is personal. Her grandmother, who raised her, was affected by dementia. "Her confused face made a deep imprint in my mind," Tsai says. "This is the biggest challenge of our lifetime, and I will give it all I have." ■

Helen Thomson *is a London-based science journalist and author of* Unthinkable: An Extraordinary Journey Through the World's Strangest Brains.

1. Iaccarino, H. F. *et al. Nature* **540,** 230–235 (2016).
2. Berger, H. *Arch. Psychiatr. Nervenkr.* **87,** 527–570 (1929).
3. Wilson, M. A. & McNaughton, B. L. *Science* **265,** 676–679 (1994).
4. Helfrich, R. F. *et al. PLoS Biol.* **12,** e1002031 (2014).
5. Koenig, T. *et al. Neurobiol. Aging* **26,** 165–171 (2005).
6. Cardin, J. A. *et al. Nature* **459,** 663–667 (2009).
7. Papalambros, N. A. *et al. Front. Hum. Neurosci.* **11,** 109 (2017).

COMMENT

DENNIS M. RIVERA PICHARDO/THE WASHINGTON POST/GETTY

The coastal city of San Juan in Puerto Rico was flooded after Hurricane Maria hit in September 2017.

Six research priorities for cities and climate change

Xuemei Bai and colleagues call for long-term, cross-disciplinary studies to reduce carbon emissions and urban risks from global warming.

Cities must address climate change. More than half of the world's population is urban, and cities emit 75% of all carbon dioxide from energy use[1]. Meeting the target of the 2015 Paris climate agreement to keep warming well below 2° C above pre-industrial levels requires staying within a 'carbon budget' and emitting no more than around 800 gigatonnes of CO_2 in total after 2017. Yet bringing the rest of the world up to the same infrastructure level as developed countries (those listed as Annex 1 to the Kyoto Protocol) by 2050 could take up to 350 gigatonnes of the remaining global carbon budget[2]. Much of this growth will be in cities in the developing world (see 'Urban development challenge').

Cities are increasingly feeling the effects of extreme weather. Many are located on floodplains, in dry areas or on coasts. In 2017, more than 1,000 people died and 45 million people lost homes, livelihoods and services when severe floods hit southeast Asian cities, including Dhaka in Bangladesh and Mumbai in India. California's suburbs and Rio de Janeiro in Brazil have experienced floods and mudslides on the heels of drought, wildfires and heavy rains. Cape Town in South Africa has endured extreme drought since 2015. By 2030, millions of people and US$4 trillion of assets will be at risk from such events (see go.nature.com/2sbj4qh).

In response, the science of cities is evolving. Urban planners and decision-makers need evidence to help them manage risks and develop strategies for climate mitigation and adaptation. Scientists are increasingly thinking of cities as complex systems and working more closely with communities. New concepts are emerging, such as smart cities.

Yet the scope and applicability of urban research is stymied: a lack of long-term studies of urban climates and their impacts makes it hard for city officials to plan decades ahead. And research grants focused on single ▶

▶ disciplines or local or national needs provide little scope for cross-disciplinary projects or comparative analyses between regions. Few online platforms exist to help cities share information and learn from one another.

Science needs to have a stronger role in urban policy and practice. Next week in Edmonton, Canada, the Intergovernmental Panel on Climate Change (IPCC) and 9 global partners will bring together some 700 researchers, policymakers and practitioners from 80 countries for the IPCC Cities and Climate Change Science Conference (https://citiesipcc.org). Participants will establish a global research agenda that will inform the IPCC special report on cities — part of the panel's seventh assessment cycle, which begins in 2023.

As members of the scientific steering committee for the IPCC conference, here we identify six priorities for cities and climate-change research.

KNOWLEDGE GAPS

Mitigating and adapting to urban climate change will require work in several areas.

Expand observations. Researchers and city authorities need to extend the quantity and types of urban data collected. The biggest gaps are in the global south. Data on informal settlements are sparse or non-existent. As well as improving availability, the coverage, quality, resolution and reliability of data need to be enhanced, and reporting should be standardized. Methodologies for remote sensing with satellites, drones and autonomous vehicles need to be developed for monitoring dense urban fabrics.

Reliable inventories of greenhouse gases are needed — from individual dwellings, factories and roads — as well as methods for verifying them. Most policymakers and practitioners still rely on city-wide or national emissions data. It is crucial to track the origins and types of air pollution with climate effects — including methane, ozone, black carbon and aerosols — because reducing these is of benefit to public health as well as to climate mitigation.

To understand the wider impacts of flooding, it is crucial to map buried networks of pipes and cables as well as hidden spaces in buildings and below ground. Researchers also need to know how people interact with infrastructure and public spaces when an extreme weather event is forecast, for example.

Narratives and local knowledge must be assimilated with technical data. Old neighbourhoods in Kano, Nigeria, have proved more resilient to floods and heat than have developments built after 1980 (ref. 3). Flat mud roofs are better at absorbing and evaporating rainwater than are metal or concrete. And Nigerian cities have historically included many open spaces, green areas and wetlands.

A global network of 'urban observatories' — diverse cities that are focal points for research and long-term monitoring — is needed. Data, research and practice should be shared online. In the United Kingdom, Newcastle University collects 1 million measurements a day from sensors across the city and makes them openly available online in real time. These range from transport emissions, precipitation, water flows and air qualities to biodiversity measures such as beehive weight. The city council and the transport, energy, environmental and water sectors are using the data. Similar observatories are being developed in Sheffield and Bristol.

To ensure trust, such data need to be verifiable and used transparently. Researchers and practitioners need to develop mechanisms for governance, security, ethics and engagement. There are privacy and security concerns; for example, many cities are now legally required to protect private data. Some are reticent to publish information that might reveal they failed to meet targets.

Understand climate interactions. Climate processes are complex — more so in cities. For example, urban air pollution in Chinese cities is causing heavier rainfall as fine particles influence clouds[4]. Impermeable surfaces, such as concrete or asphalt, hold heat and reduce evaporative cooling, amplifying urban 'heat islands'[5].

Comparative studies are needed of cities in different contexts to disentangle these interactions and to find solutions. We need to know how urban morphologies, building materials and human activities affect atmospheric circulation, heat and light radiation, urban energy and water budgets. How much permeable paving is needed to lower flood risk, as Melbourne is doing with bluestone pavements? What is the impact of reflective roads and roofs, as tried in New York and Los Angeles?

"The approaches used in cities in the global north cannot be transplanted."

Climate simulations need to account for urbanization, and be scaled down to city and neighbourhood levels. Methodologies for high-resolution risk assessment of heatwaves, coastal erosion and inundation are needed. Different approaches and models need to be compared, benchmarked and coupled with assessments of local social vulnerabilities and capabilities.

Study informal settlements. By 2050, three billion people, mostly in the global south, will be living in slums: neighbourhoods that have no mainstream governance, on land that is not zoned for development and in places that are exposed to climate-related hazards such as floods. Poor housing and basic services compound the risks for individuals and households.

Enabling these communities to adapt is a priority. Assessments are needed of grassroots efforts to address hazards. For example, community-based organizations are buying land outside flood zones and constructing resilient housing in the Philippines[6], and mapping flood risk in Gorakhpur, India. Such studies should look at formal and informal relationships and include voices from marginalized groups.

Models and analytical tools tailored to these communities need to be developed, because the approaches used in cities in the global north cannot be transplanted. Data scarcity, informal socio-economic processes and limited local capacities must be considered.

Policies need to be assessed. For example, the cities of Adama and Mekele in Ethiopia boosted affordable housing around their peripheries by increasing the area of land available for development. Keeping land tenure in government hands saves effort in the long run; large networks of infrastructure and public spaces can be planned and built coherently, without having to retrofit poorly designed areas. Informal settlements may have lessons for sustainability. Residents are often efficient at using scarce resources and reusing and recycling waste. Mitigation efforts need to support, rather than undermine, livelihoods and human well-being as well as the informal economy.

Harness disruptive technologies. The digital revolution is transforming cities. For example, urban shared-mobility schemes have improved air quality and social inclusion, and reduced congestion. In Lisbon, for instance, studies have shown that a fleet of shared taxis could maintain residents' mobility levels using only 3% of the current number of vehicles. Global adoption of shared, automated electric vehicles could cut world's vehicle stocks by one-third.

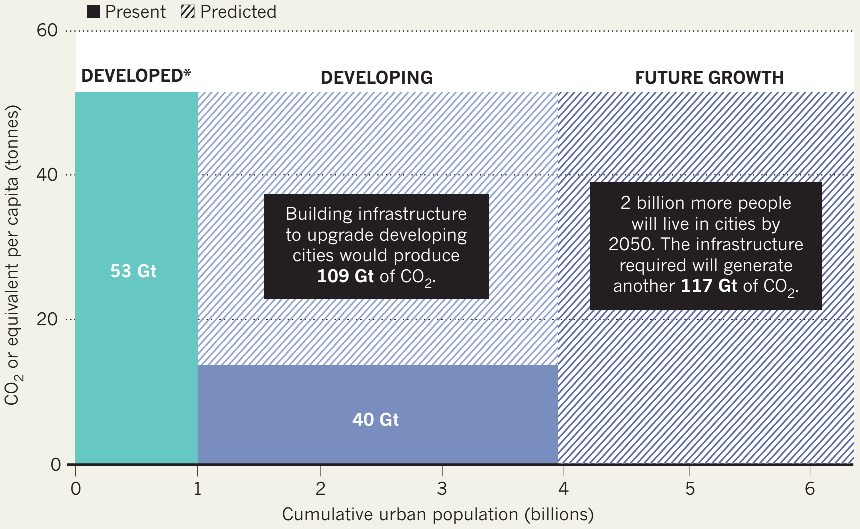
Yet these benefits could be reduced if the ease of using such technologies ultimately led to increased vehicle use. Researchers need to understand what drives positive and negative outcomes, as well as how to influence them — such as by enabling more shared travel through information and communication technologies.

Digitally connecting and controlling water, power, communications and transport systems increases the likelihood that climate hazards will affect such networks simultaneously. These coupled risks — and where responsibilities lie — are poorly understood[7]. Outages cascade and lead to shortages of food, cash and fuel, as happened in New York after Hurricane Sandy in 2012.

Affordable materials and technologies that can reduce the carbon intensity of future infrastructure in the global south should be

URBAN DEVELOPMENT CHALLENGE

Building infrastructure for fast-growing cities in developing countries could release 226 gigatonnes (Gt) of carbon dioxide by 2050 — more than four times the amount used to build existing developed-world infrastructure. To curb emissions, cities need low-carbon construction, alternative transport and better planning and design.



Building infrastructure to upgrade developing cities would produce **109 Gt** of CO_2.

2 billion more people will live in cities by 2050. The infrastructure required will generate another **117 Gt** of CO_2.

*Developed countries are as listed in Annex I to the Kyoto Protocol. Developing countries are those not listed in Annex I.

PER CAPITA DATA FROM REF. 2

developed and commercialized. For example, cement can be engineered to absorb more CO_2. Cement production is the third-largest human-made source of emissions after fossil-fuel burning and land-use change, contributing around 5.6% of global fossil-fuel and industry-related CO_2 emissions. 'Carbon-neutral' timber and bamboo have been used to build lightweight skyscrapers. These materials need to be sustainably produced at low cost. The C40's Climate Positive Development Program is experimenting in 18 cities to achieve net-negative emissions in mixed-development projects of up to 300,000 people.

Vegetation corridors, green parks, reed beds and low-lying areas that soak up water can be woven into the built environment to reduce flood and heat risks, while improving biodiversity and carbon storage. More needs to be known about the long-term performance and management of such features. Design and engineering standards need to be developed.

Support transformation. Bold strategies are needed for achieving low-carbon, resilient cities[8]. For example, China's 'sponge city' initiative helps to reduce urban flood risks by increasing green spaces, restoring wetlands and using permeable materials to absorb rainwater and delay runoff. More needs to be learned about how to change residents' lifestyles and consumption patterns, through policies and incentives, to make zero-carbon neighbourhoods and cities.

A start is to find and scale up successful local innovations. For example, Shanghai is experimenting with a range of low-carbon practices, including district heating networks, carbon labelling of consumer products and financial mechanisms[9]. Research and policy frameworks need to be developed to translate successful local innovations across cities.

Recognize global sustainability context. Cities are open, complex, dynamic systems with a global reach. Well-intended local actions can displace issues to other sectors or into the future. For example, one city's crackdown on energy-intensive production might shift the problem to less-regulated regions, with no net gain on emissions reduction. Many cities in China, South Korea and Vietnam have relocated industries outside the city to improve their environmental ratings.

A systems approach is needed to deliver on global climate change as well as the UN's New Urban Agenda and Sustainable Development Goals. More needs to be known about interactions, trade-offs and synergies between urban processes and their impacts elsewhere[10]. This entails working across disciplines and governance silos. 'Nexus approaches' that trace linkages between water, food and energy systems should be extended to other sectors, to understand the relationships between infrastructure provision, inequality and resilience, for example.

NEXT STEPS

Researchers, policymakers, practitioners and other city stakeholders need to strengthen partnerships and produce knowledge together. Universities should support data platforms and long-term research programmes in their cities, while sharing knowledge nationally and internationally. Scientists should become more engaged with policy and practice networks such as C40 Cities, ICLEI Local Governments for Sustainability and United Cities and Local Governments.

We would like to see cities establish scientific advisory boards chaired by a chief science adviser, as many government departments do. These would enhance the profile of science, build capacity and leadership, and provide a point of contact.

Funding agencies need to provide grants for cross-disciplinary research and comparative studies, especially in the global south. Cities might mandate that companies bidding for large-scale government projects in renewable energy or sustainable transport, for example, contribute money for related university research, as is required in the Australian Capital Territory. Cities should develop business models and partnerships to accelerate successful experiments and scale up ideas and technologies.

Online platforms must go beyond data sharing to help researchers, policymakers, practitioners and citizens diagnose problems, generate solutions, trial and evaluate their effectiveness and embed learning. Future Earth's Urban Knowledge-Action Network shares this ambition, but requires stable financial and institutional support to support cities globally in the long term.

Research and innovation for mitigating urban climate change and adapting to it must be supported at a scale that is commensurate with the magnitude of the problem. ■

Xuemei Bai *is professor of urban environment and human ecology at the Fenner School of Environment and Society, Australian National University, Canberra, Australia.* **Richard J. Dawson, Diana Ürge-Vorsatz, Gian C. Delgado, Aliyu Salisu Barau, Shobhakar Dhakal, David Dodman, Lykke Leonardsen, Valerie Masson-Delmotte, Debra Roberts, Seth Schultz.**
e-mail: xuemei.bai@anu.edu.au

1. Intergovernmental Panel on Climate Change. *Climate Change 2014: Mitigation of Climate Change* (eds Edenhofer, O. *et al.*) (Cambridge Univ. Press, 2015).
2. Müeller, D. B. *et al. Environ. Sci. Technol.* **47,** 11739–11746 (2013).
3. Barau, A. S., Maconachie, R., Ludin, A. N. M. & Abdulhamid, A. *Land Use Policy* **42,** 307–317 (2015).
4. Fan, J. *et al. Science* **359**, 411–418 (2018).
5. Peng, S. *et al. Environ. Sci. Technol.* 46, 696–703 (2012).
6. Dodman, D., Mitlin, D. & Rayos Co, J. *Int. Dev. Plan. Rev.* http://dx.doi.org/10.3828/idpr.2009.10 (2010).
7. Dawson, R. J. *et al. Phil. Trans. R. Soc. A* http://dx.doi.org/10.1098/rsta.2017.0298 (2018).
8. Ürge-Vorsatz, D. *et al. Nature Clim. Change* http://dx.doi.org/10.1038/s41558-018-0100-6 (2018).
9. Peng, Y. & Bai, X. *J. Clean. Prod.* **174,** 201–212 (2018).
10. Bai, X. *et al. Curr. Opin. Environ. Sustain.* **23,** 69–78 (2016).

The Nightmare (1781) by Henry Fuseli: such hallucinations can plague people with narcolepsy.

UNIVERSAL IMAGES GROUP/GETTY

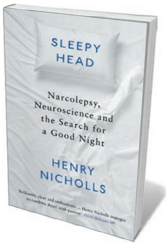
NEUROSCIENCE

The wild frontiers of slumber

Emmanuel Mignot praises a searing account of a science writer's lifelong struggle with a sleep disorder.

A seven-year-old begins to fall asleep abruptly in class, gain weight and behave aggressively. A 50-year-old man sleeps poorly and dreams vividly; suddenly, he realizes that he has had an unrecognized sleep disorder for decades. These are just some of the ways that narcolepsy can manifest. The condition involves sleepiness and abnormal rapid eye movement (REM) sleep, and affects 4 million people globally. It can remain undiagnosed and untreated for years.

Henry Nicholls, a seasoned science writer (often for these pages), has narcolepsy. Thus, his book *Sleepyhead* offers a welcome departure from most studies on the neuroscience of sleep and sleep disorders: his lived experience proves a valuable way in. His quest for the grail of perfect sleep — and the origins of his puzzling symptoms — turns into a broader journey into the complex, mysterious landscape of sleep. He meets scientists and physicians specializing in sleep disorders, and draws on historical accounts to eloquently describe conditions such as sleep paralysis, lucid dreaming, sleep apnoea and insomnia. As he shows, solutions to these harrowing problems are finally becoming available.



SLEEPY HEAD

Narcolepsy, Neuroscience and the Search for a Good Night

HENRY NICHOLLS

Sleepyhead: Narcolepsy, Neuroscience and the Search for a Good Night
HENRY NICHOLLS
Profile: 2018.

Sleep science is a young field. Its official birth is generally considered to be 1953, when REM sleep — in which the brain is active and dreaming, the body is paralysed and the eyes move rapidly — was first described by Eugene Aserinsky and Nathaniel Kleitman. Just ten years later, sleep apnoea was officially described.

Exploring concepts, diseases or symptoms chapter by chapter, Nicholls reveals that, for centuries, society and scientists dismissed sleep as a loss of time, and thus a waste of effort to study. Research into circadian rhythms has uncovered the fundamental basis of time-keeping: a gene–protein feedback loop that is present in all cells, essential to survival and reproduction. But, by contrast, as Nicholls shows, basic research has yet to reveal the molecular mechanisms that explain why we feel increasingly overwhelmed by sleepiness when we don't sleep enough (sleep debt) — although theories abound. Meanwhile, clinical approaches have exploded, leading to an oversized clinical field with relatively little basic science.

Yet the biology that does exist has immediate applications. As Nicholls shows, behaviours such as midday napping and midnight insomnia are easily explained by the interplay between circadian control of sleep and sleep-debt accumulation and release. This understanding has emerged in practical approaches such as the use of light, melatonin and sleep restriction for the treatment of insomnia.

Nicholls does much to dispel misconceptions of narcolepsy. Aside from overwhelming sleep attacks, the disease is characterized by disturbed sleep with dream-like hallucinations and sleep paralysis. Cataplexy — sudden episodes of muscle weakness typically triggered by emotions such as mirth — is another symptom that, with hynogogic hallucinations (on falling asleep), is thought to be behind the strange experience of sleep-stage dissociation: in this, a person is conscious and awake, but in REM sleep. Nicholls poignantly describes the devastating effects of narcolepsy on personal relationships, education and work, putting into context the huge cost of misdiagnoses. Too often, the hallmarks of the condition are mistreated as depression, epilepsy or simply 'conversion disorder' — physical symptoms thought to express repressed anxiety.

His stories reminded me of my early days in sleep research, in the 1990s. Well-designed studies had shown that narcolepsy was half as prevalent as multiple sclerosis — affecting 1 in 2,000 people. Yet at neurology conferences, I constantly heard: "It is impossible, I have never seen a case." Almost three decades on, I have seen thousands of patients, many of whom have 'slept their life away' undiagnosed. Given that half of all people with the condition develop it before the age of 18, this is especially tragic for children, for whom every year is critical for development.

Yet narcolepsy's cause is remarkably simple. It was discovered in 1999 by my team and that of Mashashi Yanagisawa through genetic studies in dogs and mice. In humans, research pinned it down to the loss of around 20,000 neurons in the brain's hypothalamus containing hypocretin, a wakefulness-promoting protein. The next question is how they are lost. With data suggesting an autoimmune process following an influenza infection (in which the immune system confuses parts of the flu virus with hypocretin neurons), a full understanding of the condition might teach us much about autoimmunity in the brain. A more effective treatment will be available once hypocretin-stimulating compounds are developed that can penetrate the brain, which could happen in the next decade. These compounds help narcoleptics and many other people with unexplained sleepiness.

Nicholls drives home, too, how in denial we are about our need for sleep, and the prevalence of disorders preventing it. Sleep apnoea, which involves snoring and pauses in breathing, affects 10–20% of the population, most often men. Until the 1980s, it was largely unknown to physicians, although attested in fiction: a famous sufferer is Joe in Charles Dickens' 1837 *The Pickwick Papers*. It took decades for mainstream medicine to recognize it as a frequent cause of high blood pressure. Apnoea affects people with a narrow upper airway, often due to obesity; when they breathe in, the back of their throat collapses, disturbing sleep and reducing the oxygen they take in. The standard treatment is simple — supporting the airway during sleep with pressurized air — but many find it hard to tolerate. Nothing better is available.

Some harbour fears linked to sleep. As Nicholls notes, many people with insomnia (which affects 10% of people, most often women) are terrified by their lack of sleep, and try to force themselves into it. They might spend too long in bed, making their sleep worse by reducing their sleep debt too much. This causes a vicious circle. Their anxiety might be exacerbated by the barrage of media stories on the need to sleep for eight uninterrupted hours a night, or even by accounts of fatal familial insomnia. In that extremely rare condition, slumber is impossible because of lesions in the thalamus, the brain region that filters out sensory perceptions as we fall asleep. The most effective treatment for insomnia is a better understanding of sleep physiology, notably restricting sleep to increase sleep pressure and break the vicious circle.

> ***"We are in denial about our need for sleep, and the prevalence of disorders preventing it."***

Parasomnias occupy a chunk of the book — 'almost normal' disorders such as sleep walking, night terrors and sleep paralysis. These mixed states have definitive physiological explanations. Sleepwalking and night terrors, in which people, typically children, arise screaming while still asleep, come from non-REM sleep, when one part of the brain attempts to wake the person while the cortex is still asleep. Sleepwalkers can experience fatal falls, or even have unconscious sex, which can have medical and legal consequences. In REM sleep behaviour disorder — in which REM sleep paralysis does not work and patients enact their dream — people might attack their bed partner. Most go on to develop Parkinson's disease. Research has only recently started to outline the neural underpinnings of these pathologies.

For all its strengths, Nicholls's fascinating book leaves us wanting more. As *Sleepyhead* shows, sleep sciences are still in their infancy, and current research is mostly descriptive. Luckily — although Nicholls doesn't cover this — the field is now poised to benefit from two scientific transformations. The first is genomics, which has cracked open the molecular basis of some objectively measurable traits or behaviours. The second is tools such as activity trackers, electroencephalography electrodes, devices to track movement caused by the heartbeat, and snoring recorders, which are making the objective tracking of our waking and sleeping lives vastly easier. ■

Emmanuel Mignot *is the Craig Reynolds Professor of Sleep Medicine in the Department of Psychiatry and Behavioral Sciences at Stanford University in California and the Director of the Stanford Center for Sleep Sciences and Medicine.*
e-mail: mignot@stanford.edu

FILM

Bollywood takes on menstrual stigma

Subhra Priyadarshini lauds a biopic of an inspired Indian sanitary-pad innovator.

Frugal innovation is a new norm in India, emerging sporadically in pockets of brilliance — from rural hamlets to technology labs. It has even spawned a word in Hindi: *jugaad*.

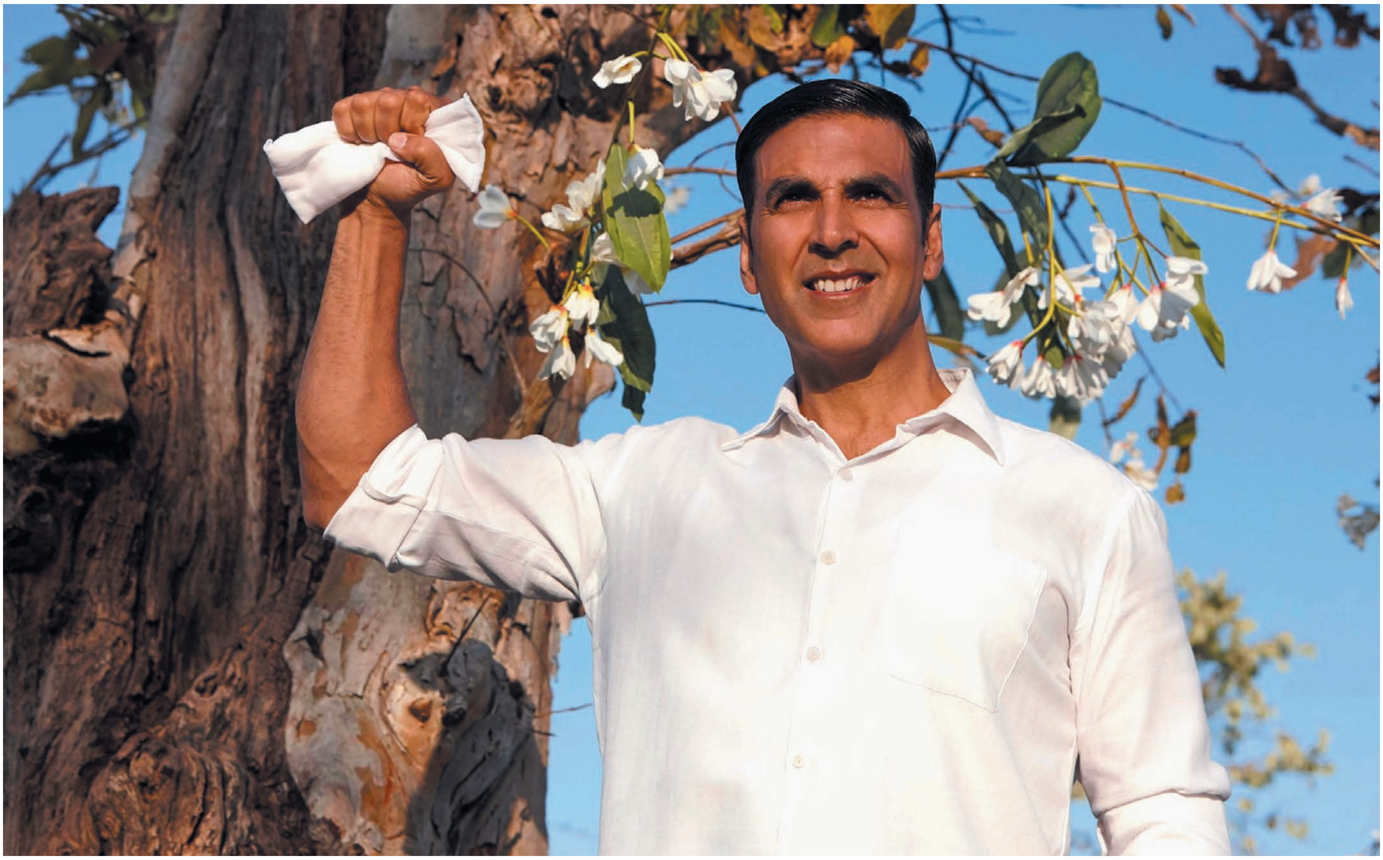
Thanks to *jugaad*, bioengineer Manu Prakash is flooding rural schools in India with his US$1 'foldscope', an origami-inspired microscope teaching science to tens of thousands of children. It is also in this spirit that, in 2000, school dropout Arunachalam Muruganantham created a do-it-yourself unit in Coimbatore, Tamil Nadu, to manufacture the world's cheapest sanitary pads. Now, Muruganantham's story hits the big screen in *Pad Man*, billed as the first feature-length film on menstrual hygiene.

'Period poverty' is a health issue affecting women in countries across the globe. In Britain, 1 in 10 girls and women aged 14–21 cannot afford sanitary products, according to London-based charity Plan International UK. In India, according to a 2015–16 government health survey, just 58% of women aged 15–24 can afford to use a hygienic method of menstrual protection: 78% in urban areas and 48% in rural ones. And the average varies wildly between states — from 91% in Tamil Nadu to just 31% in Bihar. The rest resort to rags, leaves and even ash. This can result in serious health risks, such as toxic shock syndrome, and lead to absence from school or work.

Pad Man
DIRECTOR:
R. BALKI
Columbia/Hope: 2018.

Pad Man attempts to open up this taboo topic for much-needed discussion through narrative sparked by melodrama and music. Like Shree Narayan Singh's 2017 film *Toilet: Ek Prem Katha*, centred around the problem of open defecation, it has captured the imagination of a nation grappling with a massive burden of women's health issues.

Directed by R. Balki, *Pad Man* has a starry cast. Muruganantham (renamed Lakshmikant) is played by renowned Bollywood action-hero-turned-character-actor Akshay Kumar; the powerful theatre actor Radhika Apte plays his wife, Shanthi (called Gayatri). There is even a jingoistic cameo from superstar Amitabh Bachchan, who, playing himself, declaims: "India should not be seen as a country of one billion people. ▶


COLLECTION CHRISTOPHEL/ALAMY

Akshay Kumar plays the lead in *Pad Man*, a fictionalized biopic of sanitary-pad innovator Arunachalam Muruganantham.

▸ India should be seen as a country of one billion minds."

Muruganantham's is the inspiring story of an unconventional and tenderhearted man. In the early 1990s, he was an assistant in a hardware workshop. His wife's use of rags during menstruation concerned him, so he experimented with materials — first cotton, then cellulose fibre — to make a pad that wouldn't leak, in a process of reverse engineering. At first, when it came to testing his prototypes, "the only available victim was my wife", Muruganantham said in a 2012 TED talk. In *Pad Man*, we see Lakshmikant perfecting the scientific steps of pulverizing cellulose fibres, compressing them, sealing the pad with non-woven fabric and sanitizing the whole with ultraviolet light. And it was all accomplished using four ingenious, makeshift machines that cost peanuts, compared to the giant assembly lines used by multinational companies.

At the cost of being ostracized for openly tackling a hidden issue, he worked doggedly on the pad's design. The biggest challenge was finding volunteers to test it. "Everyone thought I had gone mad," he says. He finally realized that he could turn guinea pig himself. He wore a pad, and used a deflated football filled with goat's blood and fitted with a tube. It took him six years to isolate cellulose as the core adsorbing medium. That roller-coaster journey won him a national innovation prize, a spot on *TIME* magazine's '100 most influential people' list in 2014, and one of India's highest civilian awards, the Padma Shri, in 2016.

> *"Period poverty is a health issue affecting women in countries across the globe."*

Pad Man makes Muruganantham's unusual journey relatable, although it often descends a little into preachiness. It falters, too, with a laboured first half, in which the risk to women's health is not clearly delineated, and the stigma associated with the subject of menstruation is signalled by bursts of "*Sharam*!" (shame) from the female actors. Endorsing Bollywood's unapologetic love affair with song and dance, the demure Gayatri suddenly breaks into an exaggerated, hip-swaying number to celebrate the puberty of a girl next door.

The rest of the film brings in the usual elements of a potboiler — a love angle, the rise and rise of the protagonist and Gayatri's forgive-and-feel-proud reconciliation. Bachchan delivers the applause-inducing line: "America has Superman, Spiderman and Batman. India has Pad Man!"

Over the past decade, Muruganantham has travelled across villages in India — first selling his sanitary pads, and then setting up self-sustaining pad-making units in collaboration with women's self-help groups and cooperatives. He has spawned close to 2,500 such centres, in India and a dozen other developing countries. His pads retail at a fraction of the cost of those from multinational brands.

Although *Pad Man* captures the essence of grass-roots innovation and benefits from true-to-life portrayals by a brilliant set of actors, the instructive overtone mars the narrative. If it wants to reach other countries affected by period poverty, the song-and-dance might be a dampener. The film is also currently an urban sensation. Reaching its target audience in India's rural hinterland might be difficult, given the taboo — unless Balki and team have a plan for that.

Meanwhile, a social movement is now associated with the film. Muruganantham has mentored a biologist, Maya Vishwakarma, who came home to rural Madhya Pradesh from California four years ago to spread menstrual-hygiene awareness. Vishwakarma has now received backing to distribute free pads to tribal women. Her mission has earned her the sobriquet Pad Woman. The buzz created by *Pad Man* might help her small non-profit organization to get international donors and become a national movement. ■

Subhra Priyadarshini *is chief editor of* Nature India.

SCIENCE FICTION

Ursula K. Le Guin: an appreciation

Marleen S. Barr remembers a titan of science fiction deeply inspired by anthropology.

I was watching the news this January when an image of Ursula K. Le Guin flashed up on the screen. For a second, I hoped that I was about to see a celebration of her latest achievement. When the anchor announced her death, my response was visceral.

Le Guin was a colossus of literature, and of anthropological and feminist science fiction in particular. In addition to her 22 novels, she published 11 short-story collections, 5 works of criticism, 13 books for children and several volumes of poetry. Her range was enormous, from the fantasy classic *A Wizard of Earthsea* (1968) to the great explorations of genetic engineering, gender, war and environmental despoliation in works such as *The Dispossessed* (1974). Among her many awards was the US National Book Foundation's Medal for Distinguished Contribution to American Letters, presented in 2014.

These achievements sprang from the early challenges Le Guin faced. Her first five novels, written between 1951 and 1961, were rejected by publishers who deemed them inaccessible. Happily, Le Guin remained true to her voice. She focused on clashes between worlds, and how race, gender, ethnicity, sexuality and class inform those clashes. She launched anthropology into outer space. Her extraterrestrials signify that marginalized groups are not inferior; their viewpoints are 'other' only in relation to white patriarchal hegemony.

To understand Le Guin's preoccupation with otherness, it is key to recognize that she, like many of science fiction's great literary lionesses, experienced separation from Western society in her childhood. Doris Lessing grew up in what is now Zimbabwe; James Tiptree Jr (the pen name of Alice Sheldon) travelled with her family in central Africa; Margaret Atwood spent much of her childhood in the backwoods of northern Quebec, Canada, with her entomologist father. Le Guin's separateness arose from periodic childhood immersion in the semi-wilds of California, and from her anthropologist parents.

Le Guin was born in 1929 to Alfred Kroeber, who worked with the Native Americans of California, and fellow anthropologist Theodora Kroeber. Theodora's acclaimed 1961 study *Ishi In Two Worlds* had a huge influence on her daughter. It was derived largely from the work Alfred had done (aspects of which were controversial) in the 1910s with Ishi, whose people, the Yahi, had been wiped out by genocide. The Kroeber family spent summers at the Kishamish ranch in the idyllic Napa Valley, in an intellectual milieu embracing Native American and European friends. Physicist J. Robert Oppenheimer was also a visitor.

MARIAN WOOD KOLISCH

Ursula K. Le Guin used her work to speak out about marginalized groups.

I believe that much of Le Guin's astounding ability to think outside the Western cultural patriarchal box derives from her early exposure to Ishi's story. That history informs the interplanetary cultural constructs in Le Guin's great works of the 1960s and 1970s: *Planet of Exile, City of Illusions, The Word for World Is Forest* and *The Dispossessed*. And in the 1969 *The Left Hand of Darkness*, anthropologist-emissary Genly Ai experiences an alternative to gender difference so convincing that his observation "The King was pregnant" does not seem dissonant.

Ishi, as the last of his people, might have inspired Le Guin to write her 1973 short story 'The Ones Who Walk Away from Omelas', which describes a utopia pivoting on the misery of a single abused child. The Native Americans she knew in her youth, who revered nature and did not embrace Western technology, could also have inspired the "churten principle", or "transilience" — spaceflight facilitated by storytelling — in Le Guin's 1990 short work, 'The Shobies' Story'.

Le Guin was loved by the science-fiction community because, to use the parlance of the moment, she rescued the genre from being viewed as a "shithole country". When she asked questions such as "Why are Americans afraid of dragons?" (in the eponymous essay of 1974), hers was the loudest roar emanating from science fiction's pride of literary lions. This is why, in 1994, I called Le Guin the Virginia Woolf of our time (*Foundation* **60** (spring), 58–67; 1994).

Le Guin, like Woolf, was a principled feminist with a unique, transcendent voice, in pursuit of the substrate of truth. In that 1974 essay, she defended "the uses of the imagination ... most especially in fairy tale, legend, fantasy, science fiction, and the rest of the lunatic fringe" thus:

> I believe that all the best faculties of a mature human being exist in the child, and that if these faculties are encouraged in youth they will act well and wisely in the adult, but if they are repressed and denied in the child they will stunt and cripple the adult personality. And finally, I believe that one of the most deeply human, and humane, of these faculties is the power of imagination. ■

Marleen S. Barr *teaches English at the City University of New York. She has won the Science Fiction Research Association's Pilgrim Award for lifetime achievement in science fiction criticism. Her books include* Feminist Fabulation *and* Genre Fission. *Her most recent publication is a short-fiction collection,* When Trump Changed.
Twitter: @marleenbarr

Correspondence

Italy's policy shift on immunization

Italy's experience of mandatory vaccination could complement the latest French case study you discuss (see *Nature* **553,** 249–250; 2018). It provides insight into why such a law was enforced in Italy, whether there might have been better alternatives and whether the law is working.

In 2017, it became compulsory in Italy to vaccinate infants against ten diseases: *Haemophilus influenzae* type b, measles, mumps, rubella, varicella and whooping cough (pertussis), as well as those that were already mandated (diphtheria, tetanus, polio and hepatitis B).

Unlike in France, immunization coverage in Italy had decreased alarmingly over the previous 5 years: a fall of 5.3% in 2011–15 for the measles vaccine, for example. Italy was subsequently ranked sixth-highest worldwide for measles cases in 2017 (it had 1,620; see go.nature.com/2o7jnwc). Vaccination was swiftly made mandatory.

Pilot schemes in the Veneto region (5 million inhabitants) showed that alternative strategies were not feasible. The schemes suspended the formerly mandated vaccinations and invested in health education to promote voluntary vaccine uptake. This led to a decline in coverage for polio vaccine in 2006–16, for example: by 5.2% in Veneto compared with 3.3% nationwide (see C. Signorelli *et al. Ann. Ist. Super. Sanità* **53,** 231–237; 2017).

The new law seems to be working. Preliminary data show that almost one-third of the previously unvaccinated children born in 2011–15 have now been immunized. Polio and measles vaccine uptake has increased by 1% and 2.9%, respectively, and by even more in selected regions (see C. Signorelli *et al. Lancet Infect. Dis.* **18,** 26–27; 2018).

As public-health representatives, we acknowledge that government action was epidemiologically justified. However, proactive intervention is still needed to enhance vaccine uptake and promote public trust.

Roberto Burioni, Anna Odone, Carlo Signorelli *University Vita-Salute San Raffaele, Milan, Italy.*
anna.odone@mail.harvard.edu

Democratize access to digital agronomy

Big data, field robotics and new sensing technology are set to revolutionize agriculture (see, for example, A. King *Nature* **544,** S21–S23; 2017). The international community will need to step in to democratize access to these advances, and modify them to suit the smallholders who comprise the majority of farmers worldwide.

Around one-quarter of the world's food is produced on farms smaller than 2 hectares, and about half on farms less than 20 hectares (M. Herrero *et al. Lancet Planet. Health* **1,** e33–e42; 2017). Such farms are restricted by limited infrastructure and a lack of money, so relatively few can afford advanced digital technologies. The majority must rely on mobile phones.

Although mobile-phone coverage has increased and cheap sensors can be deployed in the field, many smallholders have no Internet access and are unable to buy goods such as fertilizer or irrigation systems. There are shortfalls in the organization of supply chains, market access and advice for small farms. Such factors could stymie the vision of an agricultural revolution that is technology-based, inclusive and equitable.

To bridge these gaps, research institutes, governments, the private sector and agricultural-development organizations must commit to creating data-driven agronomy that is accessible to all.

Zia Mehrabi *University of British Columbia, Vancouver, Canada.*
Daniel Jimenez, Andy Jarvis *International Center for Tropical Agriculture, Cali, Colombia.*
zia.mehrabi@ubc.ca

Broaden behavioural addiction research

In our view, your call for government agencies to support research into gambling disorder (see *Nature* **553,** 379; 2018) should be extended to a wider group of behavioural addictions. This work can then inform policy and public-health initiatives.

The American Psychiatric Association formally recognizes only gambling disorder as a behavioural addiction. Yet the gaming industry as a whole was estimated at more than US$100 billion last year (go.nature.com/2egtu8n). As in the case of gambling, many jurisdictions do not have agencies that support research into gaming.

The Internet has facilitated the availability, affordability and accessibility of gaming and other behaviours such as shopping and viewing pornography. The extent to which problematic engagement in these activities represents distinct disorders warrants further research, particularly given controversies regarding which disorders constitute behavioural addictions. In this process, their associated harms, clinical relevance, theoretical underpinnings and empirical evidence must be considered.

The World Health Organization has held annual meetings since 2014 to discuss pressing needs, research agendas and policy initiatives related to Internet use, with gaming disorder being proposed as a formal diagnosis (see also go.nature.com/2etzndv). Understanding the biological, psychological and social processes underlying addictive behaviours stands to improve prevention and treatment strategies. This is crucial for young people, given the pervasiveness of digital technologies and the potential impact of such behaviours on development.

Marc N. Potenza *Yale University, New Haven, Connecticut, USA.*
Susumu Higuchi *NHO Kurihama Medical and Addiction Center, Yokosuka, Japan.*
Matthias Brand *University of Duisburg-Essen, Germany.*
marc.potenza@yale.edu

Don't conflate risk and resilience

'Risk' and 'resilience' are fundamentally different concepts that are often conflated. Yet maintaining the distinction is a policy necessity. Applying a risk-based approach to a problem that requires a resilience-based solution, or vice versa, can lead to investment in systems that do not produce the changes that stakeholders need.

Risk assessment and management consider efforts to prevent or defuse threats before they occur. Resilience assessment accepts the possibility of system failure and focuses on its recovery and adaptation. Resilience holds promise in many fields, including psychology, ecology and engineering, but can be misapplied in practice.

The US National Academy of Sciences, for example, defines resilience as "the ability to anticipate, prepare for, and adapt to changing conditions and withstand, respond to, and recover rapidly from disruptions". This meaning places risk within the definition of resilience: 'adapt' and 'recover' are resilience concepts; 'withstand' and 'respond to' are risk concepts. If experts wish to adopt a risk-based approach, they should focus on the ability to withstand and respond to threats. Likewise, for a resilience-based approach, they should focus on system recovery and adaptation in the aftermath of threats.

Igor Linkov, Benjamin D. Trump *US Army Corps of Engineers, Concord, Massachusetts, USA.*
Jeffrey Keisler *University of Massachusetts Boston, USA.*
igor.linkov@usace.army.mil

NEWS & VIEWS

GLOBAL HEALTH

Precision maps for public health

Researchers have produced high-resolution maps of childhood growth failure and educational attainment across Africa between 2000 and 2015, to assess progress and guide policy decisions in public health. SEE ARTICLES P.41 & P.48

BRIAN J. REICH & MURALI HARAN

Maps are key to revealing public-health challenges and suggesting potential solutions. Physician John Snow's map of cholera cases and public wells in London in the 1850s was an early example of this, famously revealing that cholera is spread through contaminated water[1]. Two papers[2,3] in *Nature* now provide high-resolution maps of health and education levels for children across Africa. Their maps will draw attention to areas most in need of support, and guide future interventions to where they can have the greatest impact.

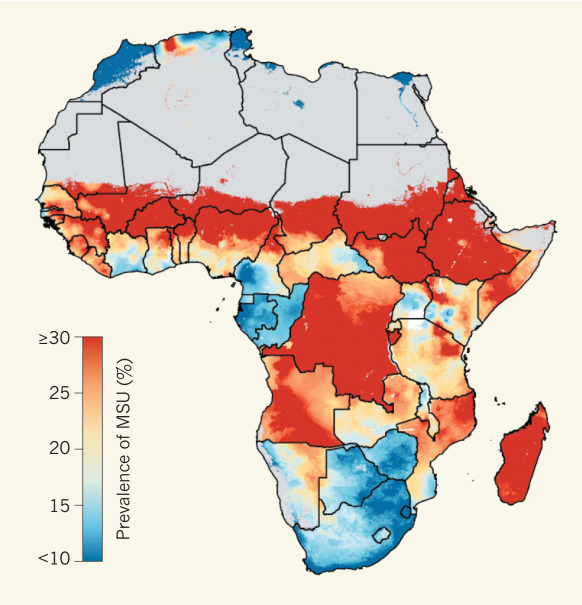


Figure 1 | Mapping public-health problems across Africa. Two studies[2,3] have produced high-resolution maps of Africa, in which each 'pixel' represents 5 × 5 kilometres. The first study documented childhood growth failure between 2000 and 2015. Here, one heat map from the study shows the percentage of children classed as moderately or severely underweight (MSU) in each pixel in 2000. Grey represents areas for which data were not collected. The second study (not shown) analysed educational attainment, and produced similar heat maps. (Map taken from Extended Data Figure 2 of ref. 2.)

Public-health policies and decision-making are often local, and so, ideally, are based on information on a small spatial scale. For example, when analysing disparities in immunization rates between urban and rural areas, information at the level of a village is often more valuable than district-level information. Like most disciplines in the twenty-first century, public health benefits from access to an unprecedented amount of data that can provide information not only on smaller spatial scales than was previously possible, but also more frequently — an exciting prospect for scientists and policymakers.

However, individual data sets often do not provide the information needed to answer key policy-relevant questions. For example, to estimate the relationship between maternal education level and infant mortality, it might be necessary to combine data from multiple sources. In much the same way as one can make educated guesses about missing pieces in a partially completed puzzle, it can be possible to find hidden information by piecing together data from different sources. But because each data source comes with its own set of errors and complexities, sophisticated statistical methods are required to both integrate the data into usable information and reflect its intrinsic uncertainties. An example of one such set of methods is Bayesian geostatistics, which combines data from multiple locations and data sets, exploiting their spatial correlations to predict values for regions for which information is lacking, while also providing information about the uncertainties involved in these predictions.

The current studies use new, advanced Bayesian geostatistical tools to analyse two problems Africa-wide. In the first paper (page 41), Osgood-Zimmerman *et al.*[2] focused on childhood growth failure. They pooled geolocated information on growth stunting, muscle wasting and weight in children under the age of 5 from several surveys across tens of thousands of villages over 15 years. They then combined this with information on local climate and geography. They carefully validated their statistical model by first fitting the model to data at a subset of locations, then comparing the predictions from this fitted model with data at a different subset of locations.

The map is split into 'pixels' of 5 × 5 kilometres across Africa, and shows changes in growth failure over time, from 2000 to 2015 (Fig. 1). The authors use their data to point out the differences in improvements across time in different regions, and show that national-level data mask nuances uncovered by their precision maps. They also provide measures of the certainty of the prediction made for each region, thereby highlighting both the uncertainty in the maps and the areas of Africa most in need of additional sampling. They then use their model to assess the likelihood of achieving 2025 global nutrition targets, and point out regions in which progress is lagging behind. The authors find that, unless there is a change in the current rates of improvement, much of the continent will fail to meet the goal of ending malnutrition by 2030.

In the second paper (page 48), Graetz *et al.*[3] used a similar approach to map local variations across Africa in the number of years of education that women between the ages of 15 and 49 have received. This is of particular interest because educational attainment is linked to the health of both mothers and their children. The authors produced their maps from data gathered from geolocated household surveys and censuses. They generated maps of average attainment across regions, including changes between 2000 and 2015, and provided uncertainties for each average. In addition, they generated similar maps for men, and and showed that, although there has been progress in educational attainment for both men and women across the continent, substantial differences between the sexes remain.

As these papers show, we now have sufficiently mature statistical methodology, theory and software to analyse continental-scale problems using sound methods and open-source software. This type of analysis was simply not possible ten years ago. In the past, researchers might have resorted to aggregating the data used in the current studies by country, at the expense of incorporating the local phenomenon that really drives the science. Alternatively, they might have restricted the analysis to one country, thus failing to exploit the full power of the data at hand. But the sophisticated geospatial tools used in the current work employ clever numerical approximations to sidestep the computational bottlenecks posed by analysing so many correlated observations. These methods are applicable to much more than just the public-health domains described here, and should provide scientific insights in many disciplines. Of course, it is important that these powerful statistical tools are not applied blindly. In both papers, the authors are careful to weight data appropriately and to validate their predictions at each step.

There is much excitement these days about the way in which enormous data sets are helping us to address many hard scientific challenges. In reality, data sets are useful only when combined with a deep understanding of the relevant science, economics or sociology, such as the impact of culture in a particular region, or details about how diseases spread. A solid understanding of how data are collected is also crucial. Rigorous scientific advances emerge when interdisciplinary teams work closely together — the current papers, which involve researchers trained in epidemiology, statistics, demography and public health, are prime examples of this.

The ultimate goal of a spatial analysis is to design interventions for maximum impact. If we understand a spatio-temporal process, we can optimize the allocation of resources in space and time. For example, consider the spread of malaria, and the effect of interventions such as bed-net distribution. A 2016 analysis[4] considered several malaria interventions, and determined the most cost-effective intervention for each 5-km^2 pixel in Africa on the basis of spatial variation in climate, mosquito populations and the current state of the disease. The results from Osgood-Zimmerman *et al.* and Graetz *et al.* should prove useful in an analogous study of optimal interventions for nutrition and education. We believe that we are entering an era in which this type of analysis can be applied broadly to improve the lives of people around the world. ■

Brian J. Reich *is in the Department of Statistics, North Carolina State University, Raleigh, North Carolina 27607, USA.*
Murali Haran *is in the Department of Statistics, Pennsylvania State University, State College, Pennsylvania 16802, USA.*
e-mails: bjreich@ncsu.edu; muh10@psu.edu

1. Johnson, S. *The Ghost Map: The Story of London's Most Terrifying Epidemic — and How It Changed Science, Cities, and the Modern World* (Riverhead, 2008).
2. Osgood-Zimmerman, A. *et al. Nature* **555,** 41–47 (2018).
3. Graetz, N. *et al. Nature* **555,** 48–53 (2018).
4. Walker, P. G. T., Griffin, J. T., Ferguson, N. M. & Ghani, A. C. *Lancet Glob. Health* **4,** 474–484 (2016).

QUANTUM COMPUTING

Qubits break the sound barrier

Quantum logic gates based on trapped ions perform more accurately than solid-state devices, but have been slower. Experiments show how trapped-ion gates can be sped up, as is needed to realize a quantum computer. SEE LETTER P.75

TOBIAS SCHAETZ

One of the goals most eagerly pursued by physicists is the development of a universal quantum computer — a machine capable of running a superposition of many correlated tasks, and one that would offer much better performance for dedicated jobs than do conventional computers. One of the most promising approaches is to use systems of trapped ions. Such systems provide the best fidelities[1,2] for all quantum logic operations (that is, they most reliably produce the correct outputs for a given input), including operations performed by the logic gates needed to process quantum bits (qubits).

Two-qubit gates are needed to implement quantum computers that can run any algorithm. In trapped-ion systems, two-qubit gates carry data between ions using phonons (quantum units of vibration) associated with the collective oscillation of the ions[3]. However, the operating speed of the gates has been limited by the oscillation frequency, which defines the speed of sound in the ensemble of ions and therefore sets the speed limit for communication. Several strategies have been proposed to overcome this limitation[4–8]. On page 75, Schäfer *et al.*[9] report the experimental realization of one such strategy, building on previously reported work in this area[7,10]. The researchers find that the speed of their two-qubit gates is more than ten times that of previously reported trapped-ion gates.

The basic principle of two-qubit gates is reminiscent of classical logic gates. Let's consider a classical controlled NOT (CNOT) gate, which turns four possible input states (00, 01, 10 or 11) into four output states (00, 01, 11 and 10, respectively). In other words, the second input bit (the target bit) is flipped between the 0 and 1 states only if the first bit (the control bit) is 1.

The quantum version of the CNOT gate allows much more than these four states to be processed. For example, the control qubit can enter the gate in a superposition state, 0 + 1. If the target qubit is 0, then running the same logic operation as for the classical CNOT flips and unflips the target at the same time. The two qubits thus end up in the final state of 00 + 11, which is a maximally entangled state: the measurement of one qubit yields a completely random output, but instantaneously fixes the state of the second qubit to be identical to that of the first, regardless of the distance between the qubits. This correlation could form the backbone of a quantum computer — it allows many operations to be run in parallel, so that a superposition of all possible inputs yields all possible results at once.

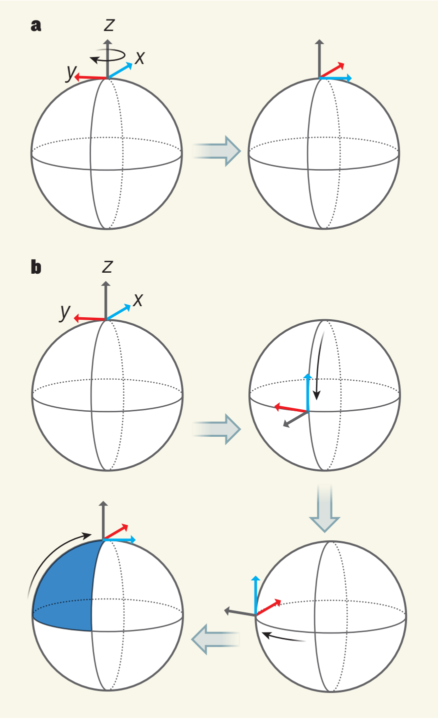


Figure 1 | Dynamical and geometric phase generation. a, When a physical state that can be described using *x*, *y* and *z* axes is placed at the north pole of a sphere, a property known as dynamical phase is generated by rotating the system around its *z* axis. The phase change equates to the change in the orientation of the *x* and *y* axes relative to their original ones. **b,** If the physical state moves down the surface of the sphere to the equator, then along the equator and back up to the north pole, the net result is rotation of the *x* axis, as in **a**. This phase change is proportional to the area enclosed by the pathway taken (blue region), and is called a geometric phase change. Schäfer *et al.*[9] report a method for generating geometric phase changes of trapped ions, and use it to implement quantum logic gates that operate much faster than similar, previously reported trapped-ion gates.

The two-qubit gate implemented by the authors can be seen as a CNOT gate, in which qubit information is encoded in quantum ‘phases’. To picture what this means, imagine an arbitrary state that can be represented by *x*, *y* and *z* axes, positioned at the north pole of a sphere (Fig. 1a). Simply rotating the system around its *z* axis produces a gain of phase: a change in the orientation of the *x* and *y* axes, which can be represented by the angle of rotation of the *x* axis within the *x*–*y* plane.

However, to implement a two-qubit gate, the phase of the gate must depend on which states are adopted by the control and target qubits, and thus must be acquired in a fundamentally different way. If the system of axes is moved down the surface of the sphere to the equator, then along the equator for a short distance and back up to the north pole, the net result is again a rotation of the *x* axis (Fig. 1b). This phase change is proportional to the area enclosed by the pathway taken, and is referred to as geometric phase.

Returning to Schäfer and colleagues’ work, the motion of a pair of ions along the pair’s axis can be described in terms of two vibrational modes — one in which the ions oscillate in sync with each other, and the second in which they oscillate in opposite directions. Oscillation in these modes correlates the ions’ positions and momenta, thus defining a ‘phase space’ for their collective motion and for geometric phase generation, analogous to the surface of the sphere mentioned above. The motion of the ions can be controlled using lasers, which induce a shift of the ions’ electronic states that depends on the position of the ions. Position-dependent shifts of states also generate a force. Because the qubit of a trapped ion is encoded by electronic states, the force exerted on the ions by the lasers depends on the state of the qubit.

If two overlapped, coherent laser beams (that is, laser beams whose light waves are in sync) are used, they produce a standing wave. If the frequency of one of the lasers is tuned slightly away from the frequency of the other (corresponding to a frequency change of δ), then the standing wave starts to move. If δ is the same as the resonance frequency of one of the ions’ vibrational modes, then the wave shakes the ions.

However, to induce geometric phase changes, δ must not be at the resonance frequency of the modes. At off-resonance frequencies of δ, the wave excites oscillations (phonons) of the trapped ions but then falls out of sync with those oscillations. After a period of time corresponding to half the amount of time needed for a gate operation, the same wave starts to decelerate the oscillation. Because the four electronic states corresponding to each of the four possible qubit combinations (00, 01, 10 or 11) couple to the lasers differently, the forces exerted on those states by the lasers are also different. The four states therefore move along different phase-generating pathways, so that the phase gain depends on the qubit combination. The phase information is then translated into qubit information by a simple operation that acts only on single qubits.

In previously reported two-qubit gates, ion displacement was performed adiabatically — the time taken for a displacement that induces geometric phase changes was long compared to the period of oscillation of the ions, limiting the operating speed of the gates[1,10]. Schäfer *et al.* have overcome this speed limit by shaping the amplitude of the laser pulses precisely in time, so that phase change is generated from pathways of a different shape from those used previously. The adiabatic situation is akin to adjusting the working of a pendulum clock by gently wobbling the clock. Schäfer and colleagues’ method is like hitting the pendulum repeatedly with well-timed hammer strikes.

Remarkably, the strikes are calibrated to work correctly no matter where in its oscillation the pendulum happens to be. This makes the logic-gate operation robust to fluctuations of and within the strikes — the fluctuations might change the pathways taken to generate geometric phase changes, but they leave the areas enclosed by the pathways unaffected. Using their method, the authors speed up their gates sufficiently to challenge the dogma that trapped-ion, two-qubit gates are slower than analogous solid-state systems, such as those that use superconducting or silicon-based qubits.

It remains to be seen whether trapped-ion qubits (or qubits based on other platforms, or combinations of qubit types) can be sufficiently well controlled and used in large enough numbers to implement a universal quantum computer. Even if all operations were to have fidelities of 99.9%, a substantial number of additional qubits would still be required for quantum-error correction; these correction processes would take up additional computational time, slowing everything down. More experiments are needed in which gates are concatenated, to find out how this concatenation affects errors.

Schäfer *et al.* suggest that optimization of the parameters needed for the lasers, ion trapping and laser–qubit coupling will enable further speed increases and improve fidelities. However, classical computers will still be needed to control the protocols performed by quantum systems, and the speed limit for any qubit platform might be set by that classical computer. It should also be noted that all operations required for quantum computers will have different speed limits. Nevertheless, speeding up quantum logic gates, while at the same time mitigating or reducing the impact of most of the disturbances that affect them, is an excellent starting point for studying how the performance of quantum devices changes as their size increases, and potentially paves the way to a quantum computer. ■

Tobias Schaetz *is at the Institute of Physics, Albert Ludwig University of Freiburg, Freiburg 79104, Germany.*
e-mail: tobias.schaetz@physik.uni-freiburg.de

1. Ballance, C. J., Harty, T. P., Linke, N. M., Sepiol, M. A. & Lucas, D. M. *Phys. Rev. Lett.* **117,** 060504 (2016).
2. Gaebler, J. P. *et al. Phys. Rev. Lett.* **117,** 060505 (2016).
3. Cirac, J. I. & Zoller, P. *Phys. Rev. Lett.* **74,** 4091 (1995).
4. García-Ripoll, J. J., Zoller, P. & Cirac, J. I. *Phys. Rev. Lett.* **91,** 157901 (2003).
5. Duan, L.-M. *Phys. Rev. Lett.* **93,** 100502 (2004).
6. García-Ripoll, J. J., Zoller, P. & Cirac, J. I. *Phys. Rev. A* **71,** 062309 (2005).
7. Steane, A. M., Imreh, G., Home, J. P. & Leibfried, D. *New J. Phys.* **16,** 053049 (2014).
8. Palmero, M., Martínez-Garaot, S., Leibfried, D., Wineland, D. J. & Muga, J. G. *Phys. Rev. A* **95,** 022328 (2017).
9. Schäfer, V. M. *et al. Nature* **555,** 75–78 (2018).
10. Leibfried, D. *et al. Nature* **422,** 412–415 (2003).

STEM CELLS

A gut feeling for cellular fate

A population of progenitor cells in the midgut of fruit flies undergoes differentiation in response to mechanical force. This finding marks the first time that such a phenomenon has been reported *in vivo*. **SEE LETTER P.103**

JACKSON LIANG & LUCY ERIN O’BRIEN

Over the past decade, advances in bioengineering have led to a newfound appreciation of the effects of mechanical force on stem cells. Micrometre-scale culture systems that can subject cells to highly specific physical deformations have allowed researchers to demonstrate that force can modulate stem-cell behaviours, and even prime stem cells for therapeutic transplantation[1,2]. However, even the most advanced culture systems merely approximate the complex and dynamic forces that stem cells experience in their native tissues. On page 103, He *et al.*[3] combine sophisticated genetic approaches and innovative physical manipulations to investigate the role of force on stem cells *in vivo*. They make the striking discovery that mechanical force drives the differentiation of a specialized population of progenitor cells in the midgut of adult fruit flies (*Drosophila melanogaster*).

The fruit-fly midgut is equivalent to the stomach and small intestine of vertebrates. All digestive organs experience physical forces that are inherent in their physiological function: ingested food distends the gut, and muscle contractions compress it. These forces continuously deform the gut's epithelial lining, which includes both mature, differentiated cells (absorptive enterocytes and hormone-secreting enteroendocrine cells) and progenitor cells (stem cells and immature daughters that are committed to, but have not yet adopted, a particular differentiated identity).

He *et al.* identified a population of progenitor cells in the fly midgut that expresses the stretch-sensitive channel Piezo — a membrane-spanning, trimeric protein complex that opens in response to mechanical stimulation to allow the passage of ions across the membrane[4]. To trace these cells *in vivo*, the authors genetically engineered the Piezo-expressing cells such that they, and any cells that arose from them, produced a fluorescent protein. This analysis revealed that the Piezo-expressing cells mature into enteroendocrine cells. He and colleagues therefore named the population enteroendocrine precursors.

The fact that enteroendocrine precursors express Piezo suggested that they might respond to mechanical stimuli. The authors tested this possibility using two approaches. First, they distended the gut tube by feeding flies a diet containing indigestible methylcellulose. Second, they compressed gut tubes *ex vivo* using a microfluidic device. It has been established[4] that mechano-activation of Piezo causes calcium ions (Ca^{2+}) to enter the cell's cytoplasm, and He *et al.* found that Ca^{2+} levels were significantly elevated in enteroendocrine precursors in both distended and compressed midguts. Crucially, Ca^{2+} levels were not elevated in distended or compressed midguts lacking the *Piezo* gene. These experiments convincingly demonstrated that the Piezo channel mediates Ca^{2+} influx in enteroendocrine precursors in response to mechanical stimuli (Fig. 1).

He *et al.* also observed that the midguts of *Piezo*-mutant flies failed to maintain normal numbers of enteroendocrine cells. This led the authors to hypothesize that Piezo-mediated Ca^{2+} influx might promote enteroendocrine differentiation. Consistent with this hypothesis, methylcellulose-distended midguts accumulated an excess of enteroendocrine cells. This effect required both Piezo and Ca^{2+} influx, and could be replicated in *Piezo* mutants by various genetic manipulations that increased cytoplasmic Ca^{2+} levels.

Taken together, these results support a scenario in which the activation of Piezo by mechanical force spurs enteroendocrine precursors to differentiate. Future investigation should reveal the physiological purpose of mechanosensitive enteroendocrine-cell production. Until then, one possibility is that a larger population of enteroendocrine cells can more-efficiently produce the myriad hormones that coordinate local and systemic responses to ingested food.

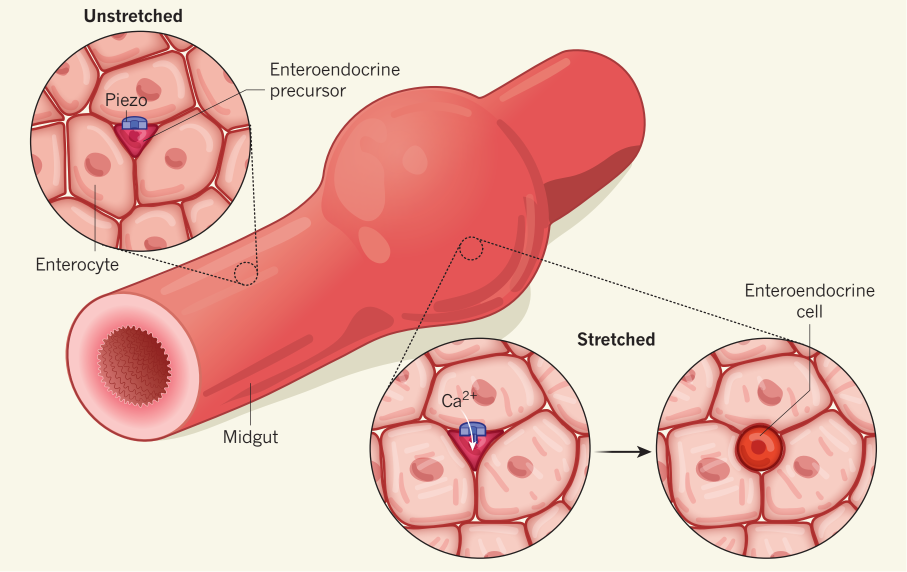


Figure 1 | Mechanosensing by specialized progenitor cells *in vivo*. The lining of the midgut of adult fruit flies contains differentiated intestinal cells, including enterocytes and enteroendocrine cells, and undifferentiated progenitor cells. He *et al.*[3] report that a subset of progenitors called enteroendocrine precursors is characterized by expression of the ion-channel protein Piezo. In unstretched conditions, the channel is closed. However, the channel opens in response to mechanical forces that stretch cells — such as gut distension from taking a meal — to allow the influx of calcium ions (Ca^{2+}). This influx promotes precursor differentiation into hormone-secreting enteroendocrine cells.

How exactly do elevated Ca^{2+} levels promote enteroendocrine differentiation? Such differentiation requires limiting the activity of a membrane-spanning receptor protein called Notch[5]. He *et al.* found that elevated Ca^{2+} levels act to inhibit Notch in enteroendocrine precursors, thus permitting their differentiation.

Interestingly, this sensitivity of Notch to Ca^{2+} might be specific to enteroendocrine precursors — the authors found that levels of Ca^{2+} had no effect on Notch activation or differentiation in enterocyte precursor cells, and others have shown[6] that the same is true of midgut stem cells. Instead, levels of Ca^{2+} in stem cells fluctuate in response to extrinsic inputs, such as nutrients, injury and stress, to control a switch between resting and proliferative states. These stark contrasts between enteroendocrine precursors, enterocyte precursors and stem cells raise intriguing questions about how different types of progenitor cell interpret the same chemical signal.

In addition, it seems that enteroendocrine differentiation is not the only cellular behaviour in the midgut to be affected by mechanical force. He *et al.* observed substantially more cell division in methylcellulose-distended midguts than in controls. However, they also found that enteroendocrine precursors rarely divide, at least under normal circumstances. One explanation that could reconcile these findings is that enteroendocrine precursors divide specifically in response to force. Another is that other, yet-unidentified, mechanosensitive progenitors exist in the midgut, and that they use a different mechanism to sense force and divide.

The identification of enteroendocrine precursors touches on a broader theme in stem-cell biology: the existence of progenitors that are specialized for specific stimuli. Perhaps the best-understood example is injury-inducible 'reserve' stem cells, which are normally in a resting state but become activated by damage[7]. Now, He and colleagues add mechanical force to the list of stimuli that are associated with specialized progenitors. Many other stimuli might also populate this list. Indeed, recent studies[8–10] using single-cell RNA sequencing have found disarming diversity among mouse intestinal stem cells, at least at the level of gene expression. This diversity might hint that we have seen just the tip of the iceberg in terms of progenitor specialization.

Are there mechano-responsive progenitors in other organs? Three mammalian organs — the intestine, lungs and skeletal muscle — would be attractive places to look. Like the fly midgut, these organs are both supported by progenitor cells and regularly subject to mechanical forces. A first step could simply be to examine them for progenitors that express Piezo or other mechanosensory channels.

Moving forward, a lack of microscale tools and protocols to manipulate force in adult tissues *in vivo* is likely to prove a bottleneck to progress. Such customized manipulations, which were crucial to the work of He and colleagues, are not easily translated between different organ systems. Perhaps the growing sophistication of three-dimensional culture systems and organ-on-a-chip devices will, over time, aid the development of technologies for mechanical manipulation of adult organs. More advances in this exciting area of research will no doubt reveal fundamental aspects of adult organ maintenance and improve strategies for tissue engineering. ■

Jackson Liang *and* **Lucy Erin O'Brien** *are in the Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA.*
e-mail: lucye@stanford.edu

1. Kumar, A., Placone, J. K. & Engler, A. J. *Development* **144,** 4261–4270 (2017).
2. Vining, K. H. & Mooney, D. J. *Nature Rev. Mol. Cell Biol.* **18,** 728–742 (2017).
3. He, L., Si, G., Huang, J., Samuel, A. D. T. & Perrimon, N. *Nature* **555,** 103–106 (2018).
4. Wu, J., Lewis, A. H. & Grandl, J. *Trends Biochem. Sci.* **42,** 57–71 (2017).
5. Sallé, J. *et al. EMBO J.* **36,** 1928–1945 (2017).
6. Deng, H., Gerencser, A. A. & Jasper, H. *Nature* **528,** 212–217 (2015).
7. Li, L. & Clevers, H. *Science* **327,** 542–545 (2010).
8. Yan, K. S. *et al. Cell Stem Cell* **21,** 78–90 (2017).
9. Haber, A. L. *et al. Nature* **551,** 333–339 (2017).
10. Barriga, F. M. *et al. Cell Stem Cell* **20,** 801–816 (2017).

This article was published online on 7 February 2018.

CONDENSED-MATTER PHYSICS

Quantum upside-down cake

Exotic states of matter called topological superconductors have potential applications in quantum computing, but have been difficult to produce in more than one dimension. A way of overcoming this limitation has now been found.

CHIH-KANG SHIH & ALLAN H. MACDONALD

Over the past decade, there has been explosive growth in the study of condensed-matter systems that have striking properties associated with unusual topologies[1]. The new kid on the block, experimentally speaking, is the topological superconductor. Like all superconductors, a topological one can transfer electric current without energy dissipation. But it also displays exotic excitations called Majorana modes, which are being investigated for use in quantum computing because they are extremely resistant to external interference[2]. One-dimensional topological superconductors have previously been reported[3,4]. Writing in *Nature Communications*, Ménard *et al.*[5] describe a technique for generating such superconductors in two dimensions.

Topological superconductors (TSCs) can be produced by combining three ingredients: ordinary superconductivity; strong coupling between the spins (magnetic moments) of electrons and their orbital motion; and the breaking of a fundamental symmetry of nature called time-reversal symmetry. The first-reported 1D TSC was realized using an ordinary superconductor, a nanoscale semiconducting wire that had strong spin–orbit coupling, and applied external magnetic fields[3]. Another 1D TSC was subsequently generated by placing ordered chains of magnetic atoms on the surface of superconducting lead, which has strong spin–orbit coupling[4]. In both cases, signatures of localized Majorana modes were observed at the ends of the TSC.

Unlike these 1D examples, a 2D TSC supports Majorana modes that are free to propagate[1], a potentially desirable feature for quantum-computing applications. Such modes travel along the edge of the TSC and are chiral, meaning that they move in just one direction. This chiral property produces a characteristic signature when the system is described in momentum space — a theoretical space that is mathematically equivalent to real space, but which offers an alternative perspective of quantum systems by replacing positions with momenta. One method for producing a 2D TSC is to couple the excitations that exist on the surface of a 3D topological insulator to an ordinary 2D superconductor[6,7]. An alternative approach is to link a 2D superconductor that has strong spin–orbit coupling to a 2D magnetic system[8].

Ménard and colleagues' work is probably the first successful realization of a 2D TSC based on the latter approach. The authors produced a 2D magnetic–superconductor composite system using a method known as epitaxy, whereby materials are grown on top of a substrate layer by layer. In this method, magnetic materials naturally separate into an array of isolated patches known as islands. The authors' experimental set-up consisted of a silicon substrate, an atomically thin film of superconducting lead, and magnetic islands made of cobalt.

Although ultrathin lead films are routinely grown on silicon substrates with atomic precision[9], the growth of magnetic islands on top of such films is technically challenging because the structural integrity of the film is often disrupted. Ménard *et al.* turned this strategy upside down: they grew their composite system in such a way that the magnetic islands formed underneath the superconducting-lead film (Fig. 1a).

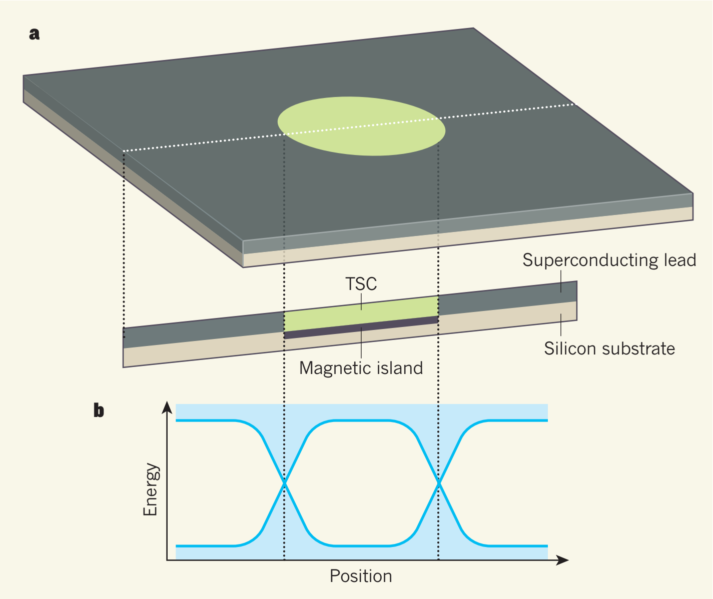


Figure 1 | Generation of a 2D topological superconductor. Ménard *et al.*[5] report a technique for producing exotic states of matter called topological superconductors (TSCs) in two dimensions. Such materials exhibit superconductivity, whereby electric current can be transferred without energy dissipation, and show excitations called Majorana edge modes, which could be useful in quantum computing. **a**, Ménard and colleagues' experiment consists of a silicon substrate, structures called magnetic islands (for simplicity, a single island is shown here) and an atomically thin layer of superconducting lead. The authors report the generation of a TSC in the region of the lead directly above a magnetic island. **b**, Ménard *et al.* obtain an energy spectrum for their system that shows the possible energies that electrons can have (light blue) at different positions. They detect X-shaped features in the spectrum at positions corresponding to magnetic-island boundaries that they interpret as evidence of Majorana edge modes.

The authors knew that if a 2D TSC could be produced, it would be located in a region of the lead film directly above a magnetic island. The growth procedure would produce seamless lateral interfaces between the 2D TSC and the surrounding lead, and Majorana modes would be expected to propagate along these interfaces.

To look for evidence of such edge modes, the authors used a scanning tunnelling microscope to obtain an energy spectrum of electrons in the lead film across the putative interfaces. The microscope had a superconducting tip to maximize the spectral resolution. However, because scanning tunnelling microscopy is a tool operating in real space, it does not directly reveal the chiral nature of Majorana edge modes, which is evident only when the energy spectrum is measured in momentum space.

In the absence of exotic excitations, the real-space energy spectrum of a superconductor contains a gap — a range of energy values that electrons cannot have. Ménard *et al.* found that gaps in the material's energy spectrum were missing at positions corresponding to magnetic-island boundaries (Fig. 1b). The authors interpreted these features as evidence of exotic edge modes. They argue that these states are topological because they are resistant to the relatively strong disorder in the lead film. Furthermore, although non-topological states could give rise to a reduced gap, they would not necessarily remove the gap entirely. The observed X-shaped gap boundaries are therefore indicative of Majorana edge modes.

Ménard and colleagues' work provides strong, albeit indirect, evidence for 2D topological superconductivity in a 2D magnetic–superconductor composite system. The evidence is indirect because it cannot rule out the possibility that non-chiral edge modes are responsible for the observed X-shaped gap boundaries. If the chiral character of the edge modes is confirmed, the authors' system might be an excellent platform for studying Majorana states, which are of interest, not only for topological quantum computation, but also in elementary-particle physics[10]. ■

Chih-Kang Shih *and* **Allan H. MacDonald** *are in the Department of Physics, University of Texas at Austin, Austin, Texas 78712, USA.*
e-mails: shih@physics.utexas.edu; macdpc@physics.utexas.edu

1. Hasan, M. Z. & Kane, C. L. *Rev. Mod. Phys.* **82,** 3045–3067 (2010).
2. Nayak, C., Simon, S. H., Stern, A., Freedman, M. & Das Sarma, S. *Rev. Mod. Phys.* **80,** 1083–1159 (2008).
3. Mourik, V. *et al. Science* **336,** 1003–1007 (2012).
4. Nadj-Perge, S. *et al. Science* **346,** 602–607 (2014).
5. Ménard, G. C. *et al. Nature Commun.* **8,** 2040 (2017).
6. Fu, L. & Kane, C. L. *Phys. Rev. Lett.* **100,** 096407 (2008).
7. Xu, J.-P. *Phys. Rev. Lett.* **112,** 217001 (2014).
8. Li, J. *et al. Nature Commun.* **7,** 12297 (2016).
9. Nam, H. *et al. Proc. Natl Acad. Sci. USA* **113,** 10513–10517 (2016).
10. Elliott, S. R. & Franz, M. *Rev. Mod. Phys.* **87,** 137–163 (2015).

This article was published online on 26 February 2018.

BIOPHYSICS

Tight complexes from disordered proteins

Charged groups on protein surfaces often take part in molecular interactions. Two unstructured proteins have been found to use charge complementarity to form a tight complex that has biologically useful kinetic properties. SEE ARTICLE P.61

REBECCA B. BERLOW & PETER E. WRIGHT

The axiom 'opposites attract' applies to many aspects of life, including the many positively and negatively charged biological molecules that control the intricate cellular processes that enable an organism to survive. Even in the crowded environment of a cell, proteins can seek out their binding partners by using charged regions to attract oppositely charged molecules. This is certainly the case for the extreme example described by Borgia *et al.*[1] on page 61, in which a high degree of opposing charge in two binding partners that lack a defined 3D structure enables the partners to associate rapidly to form a tight complex, without the need for specific interactions between amino-acid residues. Remarkably, the complex forms without either binding partner adopting a defined structure, thereby revealing a previously unknown interaction mechanism for biological molecules.

The authors sought to characterize the binding between two highly charged proteins: negatively charged prothymosin-α (Pro-Tα) and the positively charged linker histone H1.0 (H1). Both Pro-Tα and H1 are intrinsically disordered, meaning that they do not adopt defined structures in solution, but remain flexible and accessible for binding interactions. Previous studies[2] have shown that intrinsically disordered proteins typically lose some of their native flexibility when forming complexes, either by adopting a structure of their own, or by wrapping around a folded partner.

Surprisingly, Borgia *et al.* do not observe any gain of structure for either Pro-Tα or H1 on complex formation. Using a combination of nuclear magnetic resonance (NMR) spectroscopy, single-molecule fluorescence techniques and complementary computational approaches, the authors show that both proteins remain highly flexible in the complex.

Furthermore, Pro-Tα and H1 associate with extremely high affinity at physiological salt concentrations (the dissociation constant for the complex is of the order of picomolar), even though their complex is highly disordered. Because the formation of complexes is driven by complementary charge interactions, the binding strongly depends on the salt concentration and becomes much weaker as the concentration is increased beyond the physiological range. Moreover, the authors find that amino-acid residues throughout Pro-Tα and H1 are affected similarly by binding. This implies that complex formation does not depend on the existence of specific binding sites in each of the proteins — instead, interactions are distributed widely over regions of opposite charge.

Charged amino-acid residues on the surface of globular proteins are commonly associated with binding 'hot spots' — localized

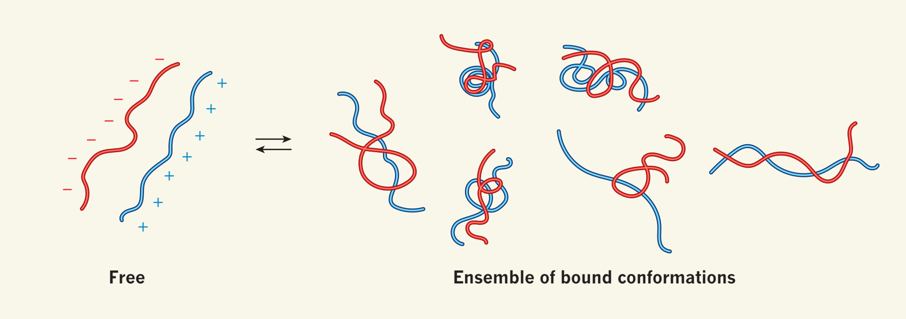


Figure 1 | Charge complementarity allows formation of tight complexes between disordered proteins. Borgia *et al.*[1] studied the formation of a complex between two proteins that, in solution, lack a defined 3D structure: prothymosin-α, which is negatively charged, and the linker histone H1.0, which is positively charged. The authors find that the complementarity of the charges on the two proteins enables them to bind reversibly and with extremely high affinity. This produces a large ensemble of bound protein conformations, many of which are adopted by only a few individual complexes and occur with approximately equal probability.

regions of the molecular surface that are involved in ligand recognition[3]. It is energetically favourable for hydrophobic amino-acid side chains in globular proteins to associate to form a hydrophobic core, a process called hydrophobic collapse. Charged amino-acid side chains are excluded from the hydrophobic core and become exposed to solvent at the surface. For disordered proteins such as Pro-Tα and H1, however, the high percentage of charged amino-acid residues precludes hydrophobic collapse, and the distribution of charges skews the conformational ensemble to a more expanded state than would be observed if the proteins were stably folded[4], leaving the charged side chains fully exposed to the solvent.

Pro-Tα and H1 become more compact when they bind to each other, probably because charge complementarity in disordered proteins can mediate the compaction of protein chains through electrostatic attractive forces, in addition to driving intermolecular interactions[5]. The authors find that the high degree of charge complementarity between Pro-Tα and H1 also provides a substantial electrostatic contribution to the binding energy of the system, making complex formation extremely favourable — as reflected by the high binding affinity of Pro-Tα for H1. The distribution of charges throughout the amino-acid sequences of both Pro-Tα and H1 allows the formation of a wide range of stable complexes that lack defined binding sites.

Many disordered proteins form ‘fuzzy’ complexes[6], which have a high degree of structural heterogeneity. Pro-Tα and H1 form an archetypal fuzzy complex that involves a large ensemble of possible bound protein conformations, many of which are adopted by only a small number of individual complexes and occur with approximately equal probability. The rate of association of Pro-Tα and H1 is limited by the diffusion of the molecules, a hallmark of electrostatic attraction[7]. The rapid association, slow dissociation and broad distribution of charge throughout the Pro-Tα and H1 sequences are responsible for the formation of the heterogeneous ensemble of complexes, in which the proteins entwine in many different configurations (Fig. 1).

The interaction mechanism of Pro-Tα and H1 probably aids their biological function. Pro-Tα assists with the assembly and disassembly of chromatin, the material in which DNA is packaged with histone proteins (such as H1) in cells[8]. To perform its function, Pro-Tα must recognize its histone substrates rapidly and with sufficient affinity to compete with the high affinity of histone–DNA interactions. The high binding affinity of Pro-Tα for H1 and the association rate of the two proteins imply that the dissociation of Pro-Tα–H1 complexes is slow enough to allow functional outcomes, but fast enough not to slow down biological turnover.

Many mechanistic questions remain. Can Pro-Tα form complexes with chromatin-bound H1, to promote the dissociation of H1 from DNA and to usher it to a new binding site, without having to wait for spontaneous dissociation? Similarly, can Pro-Tα remain bound to H1 once it has been deposited at a new DNA site? The flexibility in the Pro-Tα–H1 complexes would facilitate such processes: the positively charged regions of H1 would be exposed even when in complex with Pro-Tα, and thus be available for simultaneous binding to chromatin.

It is evident that the amino-acid sequences of Pro-Tα and H1 have a crucial role in dictating the proteins’ molecular function. The amino-acid sequences of disordered regions in proteins evolve rapidly, yet recent studies have shown that the net charge is conserved despite a high degree of sequence diversity (see ref. 9, for example). Highly charged proteins such as Pro-Tα and H1 might therefore be more tolerant to mutation than their less-charged counterparts. As noted by Borgia and colleagues, many disordered proteins have levels of net charge similar to those of Pro-Tα and H1, suggesting that the formation of dynamic complexes between disordered proteins of opposite charge might be common.

Charge complementarity between disordered proteins and their molecular partners is of great importance to signalling pathways that rely on post-translational modifications (protein modifications that occur after protein biosynthesis), and to phase-separation processes that result in the formation of concentrated droplets of proteins and nucleic acids[10]. Although it has long been evident that electrostatic interactions have a central role in the formation of biological complexes, Borgia and colleagues’ work highlights how crucial these attractive forces can be for the assembly of very strong, yet highly dynamic, molecular complexes in the cell. ■

Rebecca B. Berlow *and* **Peter E. Wright** *are in the Department of Integrative Structural and Computational Biology and The Skaggs Institute of Chemical Biology, The Scripps Research Institute, La Jolla, California 92037, USA.*
e-mail: wright@scripps.edu

1. Borgia, A. *et al. Nature* **555,** 61–66 (2018).
2. Mollica, L. *et al. Front. Mol. Biosci.* **3,** 52 (2016).
3. DeLano, W. L. *Curr. Opin. Struct. Biol.* **12,** 14–20 (2002).
4. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. *Proc. Natl Acad. Sci. USA* **107,** 8183–8188 (2010).
5. Mittag, T. *et al. Proc. Natl Acad. Sci. USA* **105,** 17772–17777 (2008).
6. Tompa, P. & Fuxreiter, M. *Trends Biochem. Sci.* **33,** 2–8 (2007).
7. Zhou, H. X. & Bates, P. A. *Curr. Opin. Struct. Biol.* **23,** 887–893 (2013).
8. George, E. M. & Brown, D. T. *FEBS Lett.* **584,** 2833–2836 (2010).
9. Zarin, T., Tsai, C. N., Nguyen Ba, A. N. & Moses, A. M. *Proc. Natl Acad. Sci. USA* **114,** E1450–E1459 (2017).
10. Wright, P. E. & Dyson, H. J. *Nature Rev. Mol. Cell Biol.* **16,** 18–29 (2015).

This article was published online on 21 February 2018.

COSMOLOGY

A surprising chill before the cosmic dawn

An experiment to estimate when stars began to form in the Universe suggests that gas temperatures just before stars appeared had fallen well below predicted limits, and that dark matter is not as shadowy as was thought. See Letter p.67

LINCOLN GREENHILL

The first stars to form generated copious fluxes of ultraviolet radiation that suffused the early Universe — a phenomenon referred to as the cosmic dawn. Many calculations have been performed to estimate when this occurred[1], but no data-driven constraints on the timing have been available. On page 67, Bowman *et al.*[2] report what might be the first detection of the thermal footprints of these stars, tracking back to 180 million years after the Big Bang.

Less than one million years after the Big Bang, the Universe consisted of atomic gas (chiefly hydrogen) and a form of matter that outweighs regular matter by more than five times[3] but has yet to be seen directly. Measurements over decades have indicated that, oddly enough, this ‘dark’ matter interacts with itself and with regular matter only through the action of gravity. It was mainly the gravity of dark matter that amplified small, localized density perturbations in the Universe shortly after the Big Bang to generate the first large-scale structures. But it was the hydrogen within these perturbations that collapsed piecemeal to form stars, bringing about the cosmic dawn.

The observable thermal footprints of early stars derive from small variations in the ratio of the number of interstellar hydrogen atoms found in two particular energy states;

a transition between these states causes a photon to be emitted or absorbed at a characteristic radio frequency. The ratio reflects the degree of excitation of the hydrogen, and can be expressed as a temperature, known as the atomic spin temperature (T_S).

At early times, when the Universe was relatively small and mean gas density was high, collisions between atoms were frequent. T_S was therefore the same as the kinetic temperature of the gas (T_G), an indicator of the energy available to excite atoms through collisions. By the time stars began to form, the Universe had expanded. Both T_G and mean gas densities had fallen, and collisions were infrequent, allowing T_S to drift upward to the temperature of the radiation (T_R) left over from the Big Bang (Fig. 1). T_R also fell as the Universe expanded, but not as quickly as T_G.

A long-standing theory[4,5] that still awaits testing predicts that absorption of UV radiation from early stars by nearby clouds of hydrogen could have driven T_S back down to T_G, but not lower. In other words, the cosmic dawn would make the gas seem colder when observed at radio frequencies. This would create an absorption feature in the spectrum of the background radiation left over from the Big Bang.

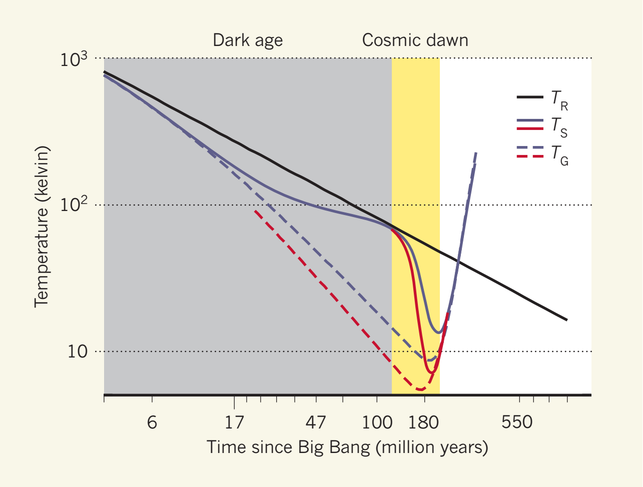


Figure 1 | Temperature changes during the evolution of the Universe. The first two phases of the Universe were the 'dark age', before stars formed (grey), and the cosmic dawn (yellow), when clouds of hydrogen in early structures collapsed to form stars. The temperature of radiation (T_R) left over from the Big Bang has declined slowly over time. The spin temperature (T_S) of hydrogen that has not formed stars reflects the excitation state of the hydrogen atoms (solid blue line shows previous estimates of T_S based on models). Bowman *et al.*[2] use observations to estimate T_S, and find that it dropped to lower values (red solid line) than predicted by models. Barkana[7] proposes that this could be evidence for a previously unrecognized, non-gravitational interaction between normal and dark matter. Such interactions would mean that the 'kinetic' temperature of gas (T_G) in the Universe also dropped to a lower minimum (red dashed line) than is predicted by known physics (blue dashed line). (Adapted from ref. 12.)

Bowman *et al.* now report the possible detection of just such an absorption signal. The authors measured T_S, averaged over much of the sky and over a contiguous range of radio frequencies; each frequency provides a window on a different time in the Universe's past. The measurement is very difficult because it must be performed using an extremely well-calibrated VHF radio antenna and receiver, to enable the weak cosmological signal to be separated from much stronger celestial signals and from those within the electronics systems of the apparatus used.

The putative absorption signal extends over a wide frequency range, one end of which looks as far back as 180 million years ago, in good agreement with theoretical predictions[6]. Remarkably, however, the peak amplitude of the absorption is two to three times larger than predicted by the most optimistic models, and the absorption profile is flat-bottomed, rather than curvilinear and Gaussian-like, which is also at odds with models.

So how can the differences from the models be explained? On page 71, Barkana[7] argues that models could achieve the reported signal amplitude and profile if non-gravitational interactions — like those that occur between charged particles — occur between dark matter and normal-matter particles, and if the dark-matter particles have relatively low masses and velocities that are less than the speed of light. The effects of variously hypothesized types of dark matter have been calculated previously[8,9], but only those in which dark matter and normal matter scatter each other increase the magnitude of the absorption signature. The idea that a detectable radio signal from the cosmic dawn can be connected to the particle properties of dark matter suggests a potentially revolutionary angle for exploring fundamental physics.

Bowman and colleagues' claim to have detected the long-sought absorption signal is bolstered by myriad tests in which the authors altered their experimental hardware or data analysis, in a concerted effort to identify systematic errors that might be responsible for the measured signal. The tests included repeating the data acquisition and analysis using a duplicate antenna at a second, nearby location; orienting the antenna at different angles with respect to the compass; and changing the ways in which the antenna is isolated from the ground. Other tests focused on switching various facets of the data calibration on and off.

However, the most stringent test will be to compare the current results with those to come from independent experiments also aimed at detecting the cosmic-dawn signal[10,11]. I hope that the unexpected amplitude and line shape of the reported absorption signal is indeed a hard-won breakthrough that reveals evidence of unexpected physics. But it is possible that systematic errors have escaped detection by the tests that were run. Two extensions to the reported tests include using circuitry that more precisely imitates the antenna than Bowman and colleagues' circuitry when attached to the receiver during performance evaluation and calibration, and the cross-checking of performance models for the antenna (which are currently based on computer simulations of antenna electromagnetics) with field measurements made when narrowband or sinusoidal signals are broadcast near the antenna.

Bowman and co-workers' report will be recognized as a milestone for this nascent experimental field: the first reputable claim of a much-anticipated detection. The follow-up will not be limited to ever finer interpretations of increasingly accurate one-dimensional spectra. Studies of the cosmic-dawn signal using interferometers (arrays of antennas) could describe the 3D structure of the Universe at that time and, by extrapolation, during the primordial 'dark age' when large-scale structure in the Universe first formed. One of Barkana's particularly notable predictions is that, if non-gravitational interactions between normal and dark matter do exist, then the absorption signal detectable by interferometers could be stronger and more distinctive than had been predicted. It would encode the spatial fluctuations of matter density that occurred during the dark age, rather than just gas temperature, thus presenting new opportunities for tests of fundamental physics. ■

Lincoln Greenhill *is in the Radio and Geoastronomy Division, Smithsonian Astrophysical Observatory, Cambridge, Massachusetts 02138, USA, and in the Department of Astronomy, Harvard University, Cambridge.*
e-mail: greenhill@cfa.harvard.edu

1. Pritchard, J. R. & Loeb, A. *Rep. Prog. Phys.* **75,** 086901–086935 (2012).
2. Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. *Nature* **555,** 67–70 (2018).
3. Ade, P. A. R. *et al. Astron. Astrophys.* **571,** A16 (2014).
4. Wouthuysen, S. A. *Astrophys. J.* **57,** 31–32 (1952).
5. Field, G. B. *Astrophys. J.* **129,** 536–550 (1959).
6. Cohen, A., Fialkov, A., Barkana, R. & Lotem, M. *Mon. Not. R. Astron. Soc.* **472,** 1915–1931 (2017).
7. Barkana, R. *Nature* **555,** 71–74 (2018).
8. Evoli, C., Mesinger, A. & Ferrara, A. *J. Cosmol. Astropart. Phys.* 024 (2014).
9. Tashiro, H., Kadota, K. & Silk, J. *Phys. Rev. D* **90,** 083522 (2014).
10. Price, D. C. *et al.* Preprint at http://arXiv:1709.09313 (2017).
11. Singh, S. *et al.* Preprint at http://arXiv:1710.01101 (2017).
12. Pritchard, J. & Loeb, A. *Phys. Rev. D* **78**, 103511-6 (2008).

ARTICLE

OPEN

doi:10.1038/nature25760

Mapping child growth failure in Africa between 2000 and 2015

Aaron Osgood-Zimmerman[1]*, Anoushka I. Millear[1]*, Rebecca W. Stubbs[1], Chloe Shields[1], Brandon V. Pickering[1], Lucas Earl[1], Nicholas Graetz[1], Damaris K. Kinyoki[1], Sarah E. Ray[1], Samir Bhatt[2], Annie J. Browne[3], Roy Burstein[1], Ewan Cameron[3], Daniel C. Casey[1], Aniruddha Deshpande[1], Nancy Fullman[1], Peter W. Gething[3], Harry S. Gibson[3], Nathaniel J. Henry[1], Mario Herrero[4], L. Kendall Krause[5], Ian D. Letourneau[1], Aubrey J. Levine[1], Patrick Y. Liu[1], Joshua Longbottom[3], Benjamin K. Mayala[1], Jonathan F. Mosser[1], Abdisalan M. Noor[6,7], David M. Pigott[1], Ellen G. Piwoz[5], Puja Rao[1], Rahul Rawat[5], Robert C. Reiner Jr[1], David L. Smith[1], Daniel J. Weiss[3], Kirsten E. Wiens[1], Ali H. Mokdad[1], Stephen S. Lim[1], Christopher J. L. Murray[1], Nicholas J. Kassebaum[1,8]§ & Simon I. Hay[1,3]§

Insufficient growth during childhood is associated with poor health outcomes and an increased risk of death. Between 2000 and 2015, nearly all African countries demonstrated improvements for children under 5 years old for stunting, wasting, and underweight, the core components of child growth failure. Here we show that striking subnational heterogeneity in levels and trends of child growth remains. If current rates of progress are sustained, many areas of Africa will meet the World Health Organization Global Targets 2025 to improve maternal, infant and young child nutrition, but high levels of growth failure will persist across the Sahel. At these rates, much, if not all of the continent will fail to meet the Sustainable Development Goal target—to end malnutrition by 2030. Geospatial estimates of child growth failure provide a baseline for measuring progress as well as a precision public health platform to target interventions to those populations with the greatest need, in order to reduce health disparities and accelerate progress.

Child undernutrition increases the risk of neonatal and child mortality and future maternal reproductive outcomes[1–3]. Child growth failure (CGF) is the specific subset of child undernutrition, excluding micronutrient deficiencies, that is characterized by the relationship between insufficient height and weight at a given age, and this subset is most universally described in terms of univariate 'growth standards', for which age-specific heights and weights are compared to healthy reference populations[4,5]. In aggregate, univariate assessments of stunting, wasting and underweight (Extended Data Fig. 1) can serve as a comprehensive assessment of CGF. Prevalence of moderate and severe stunting, wasting and underweight among children aged 0–59 months is defined as the proportion of children with a height-for-age, weight-for-height or weight-for-age *z* score that is more than two standard deviations below the 2006 WHO (World Health Organization) growth reference population, respectively[4].

The Millennium Development Goals (MDG) had a single nutrition target: a 50% reduction in prevalence of underweight in children under five between 1990 and 2015. In 2012, WHO member states endorsed a broader agenda to improve nutrition by 2025: the Global Nutrition Targets (WHO GNT), including stunting, wasting, low birth weight and overweight[6] in children under five (see Extended Data Fig. 1). Sustainable Development Goal (SDG) 2.2 is even more aspirational, calling for an end to all forms of malnutrition by 2030, progress towards which can be seen as inseparable from many of the other SDG child health ambitions[7–9].

Quantitative assessments of levels and trends in CGF indicators serve as key input to discussions of progress and areas for improvement[1,10–14]. According to findings from the Global Burden of Diseases, Injuries, and Risk Factors Study 2016 (GBD 2016), an estimated 36.6% of children under five were stunted, 8.6% wasted and 19.5% underweight in sub-Saharan Africa (SSA) in 2015[1]. Furthermore, CGF was the second leading risk factor for child mortality in SSA, accounting for more than 23% of deaths of children under five in this region[1].

Precision public health and child growth failure

Although country-level estimates are useful for international comparisons and benchmarking, they mask disparities in CGF at the lower administrative levels at which most health and nutrition policy planning and implementation occur. The value of precision public health in this context—the use of more spatially resolved data to guide efficient targeting of interventions to those populations with the greatest need—is increasingly recognized by the global health community[15]. This approach enables quantification of inequalities and identification of successes and failures of programmes and policies at the local level. Similar efforts that mapped subnational malaria prevalence, incidence, and mortality[16,17] have, when overlaid with interventions, shown where use of insecticide-treated nets or access to treatment is lacking, pinpointing where remedial actions are needed. Without comparable, robust subnational information on stunting, wasting and underweight, health authorities face sizeable challenges to precisely target and thus optimally fund relevant CGF interventions.

Subnational assessments of CGF have been conducted in select countries in Africa, including states in Nigeria[18], regions in Uganda[19], governorates in Egypt[20] and districts in Ethiopia[21], Malawi[22,23],

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington 98121, USA. [2]Department of Infectious Disease Epidemiology, Imperial College London, London SW7 2AZ, UK. [3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7FZ, UK. [4]Commonwealth Scientific and Industrial Research Organisation, St Lucia, Queensland 4067, Australia. [5]Bill & Melinda Gates Foundation, Seattle, Washington 98109, USA. [6]Kenya Medical Research Institute-Wellcome Trust Collaborative Programme, Nairobi, Kenya. [7]Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 7FZ, USA. [8]Department of Anesthesiology and Pain Medicine, Seattle Children's Hospital, Seattle, Washington 98105, USA.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

Tanzania[23] and Zambia[23], as well as the Demographic and Health Surveys, which report at the first administrative subdivision in 39 countries[24]. Although this initial work has unveiled coarse subnational disparities in CGF, it provides an incomplete picture, with heterogeneity remaining within administrative units. Model-based geostatistics, a set of statistical techniques developed to make inferences from spatially correlated phenomena, have produced high-spatial-resolution estimates of nutrition indicators in Burkina Faso[25], Ghana[25], Kenya[26], Mali[25], Nigeria[26], Tanzania[26] and Somalia[27]. These studies demonstrate that geo-referenced anthropometric survey data, if properly harnessed with spatially and temporally explicit models and appropriate covariates, can allow for the synthesis of these data into gridded maps. However, a sizable geographical knowledge gap remains, as the combined analyses of previous studies are not comprehensive or generalizable. Furthermore, advances in data sharing and computational statistics enable high-resolution estimates to be made over continental and global scales[16,17].

Here we provide a comprehensive geospatial analysis of CGF in 51 African countries from 2000 to 2015, offering highly relevant subnational information on key nutrition indicators for policymakers and health practitioners at all administrative subdivisions. We used Bayesian model-based geostatistics, which uses geo-referenced child anthropometry survey data and gridded covariates over space and time, in an ensemble modelling framework based on stacked generalizations[28] and spatial validation processes, to produce 5 × 5 km gridded estimates of stunting, wasting, and underweight for children under five. To ensure comparability with national estimates and to facilitate benchmarking, we calibrated pixel-level estimates to those produced by GBD 2016[1] and subsequently aggregated 5 × 5 km estimates to multiple administrative subdivisions in each country. We compared the annualized rate of change (AROC) for each CGF measure during the MDG era (2000 to 2015) relative to the AROC required between 2015 and 2025 to meet the WHO GNT (Figs 1g, 2g, Extended Data Fig. 2g), and the acceleration in the pace of progress required between 2015 and 2025 to achieve the WHO GNT (Figs 1i, 2i, Extended Data Fig. 2i).

Disparate progress in reducing child growth failure

Between 2000 and 2015, nearly all African countries showed a reduction in the absolute levels of stunting, wasting and underweight in children under five, but observed rates of change varied markedly[1,2,10,29]. If current rates of progress continue, many countries are on track to meet the relevant WHO GNT[6] at the national level. This includes most of eastern and southern SSA and the coastal sections of western SSA at more local scales. However, our results also show particularly high levels of CGF, with little evidence of improvement, across the Sahel.

Stunting (Extended Data Fig. 1a) was the most prevalent form of CGF across all years, and its change in prevalence across time was visually striking (Fig. 1a–c). While large areas of Algeria, Mozambique, Burkina Faso and Ghana showed a reduction in the prevalence of stunting from 2000 to 2015, progress in other countries was more spatially heterogeneous. Progress occurred between 2005 and 2015 in many areas, as illustrated by the Imo state in Nigeria, in which the mean estimated stunting prevalence was nearly halved (46.2% reduction; 95% uncertainty interval, 38.7–54.9%) from 31.5% (28.1–35.5%) in 2005 to 16.9% (14.7–19.2%) in 2015. By 2015, lower levels were found in coastal central Africa, particularly in areas within Ghana, Gabon and Equatorial Guinea. By contrast, northern Nigeria, southern Niger, Democratic Republic of the Congo (DRC), Zimbabwe and northern Mozambique all had areas with a prevalence of stunting near or above 40% in 2000, which was as high as 64.9% (59.3–70.8) in the Lubango municipality within Huila province, Angola. Although many of these regions showed improvement (the prevalence rate in Lubango dropped to 31.5% (27.2–35.9%) in 2015), some areas, such as regions of the Northern Province of Zambia, northern Nigeria, and southern Niger, had the highest prevalence rates in both 2000 and 2015.

Wasting (Extended Data Fig. 1b) is a short-term phenomenon that encompasses both moderate acute malnutrition and severe acute malnutrition[4]. Wasting is more sensitive to external environmental fluctuations, such as crop yields and food availability[30], and is most likely to affect children over the course of months, rather than years. These shorter-term events drive uneven temporal patterns of decline compared to the more consistent decreases seen in stunting and underweight. As such, some areas in northern Kenya, eastern Ethiopia, northern Nigeria and Madagascar show temporal variation and increases across the study years (Fig. 2a–c). The Afar region in Ethiopia, for example, had a high prevalence in both 2000 (16.7% (14.5–19.4%)) and 2015 (21.7% (18.9–24.7%)). While the estimated prevalence in regions of Madagascar also increased, these estimates were relatively uncertain (Fig. 2f). High prevalence of wasting appears in a band across the continent, with concentrations in Niger (19.9% (18.9–20.9%)), South Sudan (21.0% (18.0–24.2%)) and Burkina Faso (18.9% (17.9–19.9%)) in 2000. Foci of higher prevalence remain even in countries with low rates nationally. Kenya, for example, had a national prevalence of 5.7% (5.2–6.2%) in 2015, although rates as high as 28.2% (24.8–31.8%) were found in areas within the Rift Valley province. Prevalence of wasting in southern Africa, by contrast, remained consistently low across the study period. Some countries, including the DRC, experienced sizeable progress both nationally and subnationally, dropping from 14.6% (13.9–15.4%) in 2000, with rates as high as 18.4% (16.7–20.2%) in the Équateur province, to 8.7% (8.2–9.2%) in 2015, with a decrease in Équateur to 9.8% (8.7–11.0%), lessening the gap between national and subnational prevalence.

Of particular note when comparing the prevalence of underweight (Extended Data Fig. 1c) across time and space, is the persistent band of high prevalence across the Sahel, stretching from southern Mali in the west to the Horn of Africa in the east (Extended Data Fig. 2a–c). By contrast, for the northern coast of Africa, in countries near the Gulf of Guinea, and in southern Africa, prevalence remained low throughout the study period, achieving rates such as 3.8% (2.7–5.2%) in the Litoral province of Equatorial Guinea by 2015. Patterns of change in central Africa were highly spatially heterogeneous in countries, such as Nigeria, which achieved rates below 10% in some regions, and in excess of 30% in northern areas by 2015. Marked progress was seen in central Africa, with Rwanda reducing national prevalence from 22.3% (20.7–23.8%) in 2000 to 8.7% (7.8–9.6%) in 2015. Although Angola and the DRC have experienced substantial improvement, hot spots remain, such as the Kasai-Occidental province in the DRC, in which prevalence was 25.3% (22.9–27.9%) in 2015.

For each CGF metric, Figures 1e, 2e and Extended Data Figure 2e show estimates of the population-weighted highest and lowest 10% of pixels across the continent in 2000 and 2015, as well as their overlap. Overlaid stippling across the continent represents areas in the estimated maps that experienced the 10% lowest and highest rates of decline for the 16 years that were modelled. Areas in Angola (Fig. 1e) experienced some of the highest rates of stunting in 2000, but also some of the highest annualized rates of decline; by 2015, pixels in that same area were no longer in the worst 10%. Conversely, as demonstrated for southern Niger and northern Zambia, where some of the highest stunting rates in both 2000 and 2015 occurred, these maps can elucidate places and populations left behind as the continent progresses towards the WHO GNT.

Figures 1f, 2f and Extended Data Fig. 2f show our estimates contrasted with their respective certainty for each 5 × 5-km area in 2015. These maps more intuitively highlight areas for which our estimates are less uncertain, and the corresponding relative prevalence of the CGF indicator. For example, much of Zimbabwe had a low prevalence of wasting (3.8% (3.4–4.1%)) at the national level and had low uncertainties relative to other areas. Areas in Chad (such as the Kanem region, with a prevalence of stunting of 50.0% (47.0–52.9%)) had a high prevalence and were relatively certain. By contrast, the Melaky region of Madagascar experienced a high prevalence of wasting, but estimates in

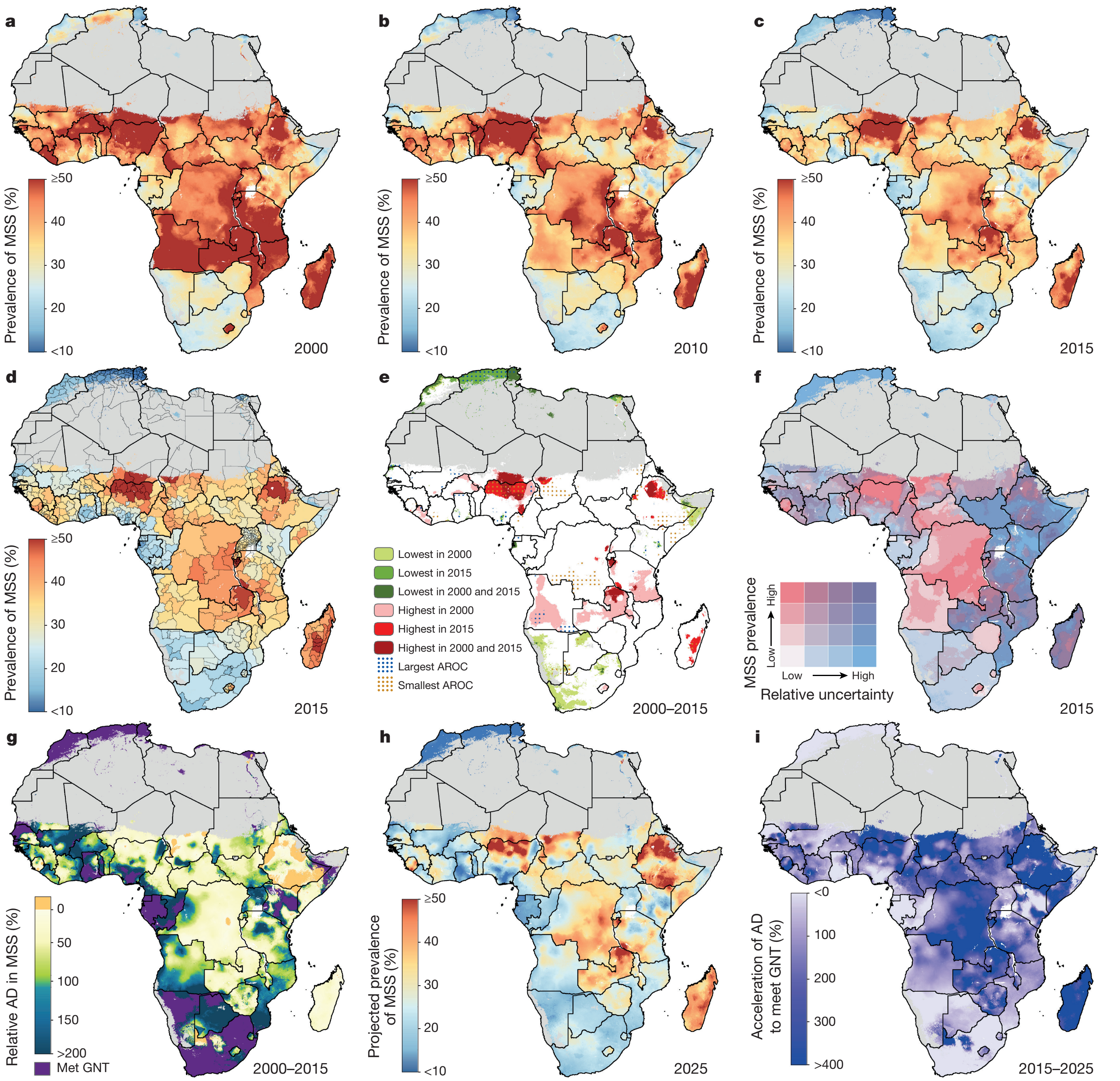


Figure 1 | Prevalence of stunting (2000–2015) in children under five and progress towards 2025. a–c, Prevalence of moderate and severe stunting (MSS) at the 5 × 5-km resolution in 2000 (**a**), 2010 (**b**) and 2015 (**c**). **d**, Prevalence of stunting at the first administrative subdivision in 2015. **e**, Overlapping population-weighted lowest and highest 10% of pixels and AROC in stunting from 2000 to 2015 across the continent. **f**, Overlapping population-weighted quartiles of stunting and relative 95% uncertainty in 2015. **g**, Annualized decrease (AD) in stunting prevalence from 2000 to 2015 relative to rates needed during 2015–2025 to meet the WHO GNT. 100% indicates the annualized decrease from 2000 to 2015 equivalent to the pace of progress required during 2015–2025 to meet the WHO GNT by 2025 (40% decrease in stunting, relative to 2010). Blue pixels exceeded this pace; green to yellow pixels proceeded at a slower rate than required; orange pixels were non-decreasing; and purple pixels were estimated to have met the target by 2015 ('Met GNT'). **h**, Pixel-level prevalence of stunting was predicted for 2025 on the basis of the annualized decrease achieved from 2000 to 2015 and projected from 2015. **i**, Acceleration in the annualized decrease in stunting required to meet the WHO GNT by 2025. Purple pixels were either non-decreasing or must accelerate their rate of decline by more than 400% over 2000–2015 rates during 2015–2025 to achieve the target; white pixels require no increase. Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].

those areas were relatively uncertain (42.2% (28.3–58.0%)). For more detail, see Supplementary Figs 13–15.

The predicted space–time models of CGF prevalence closely matched the observed national survey data, and we used 5-fold cross-validation strategies to assess the fit of our models. The full array of validation metrics by indicator and country are provided in the Supplementary Information (Supplementary Tables 8–19 and Supplementary Figs 16–36).

Given the continental scope and fine spatial scale of this work, additional results are provided in the Supplementary Information and all outputs of these analyses at the first administrative subdivision

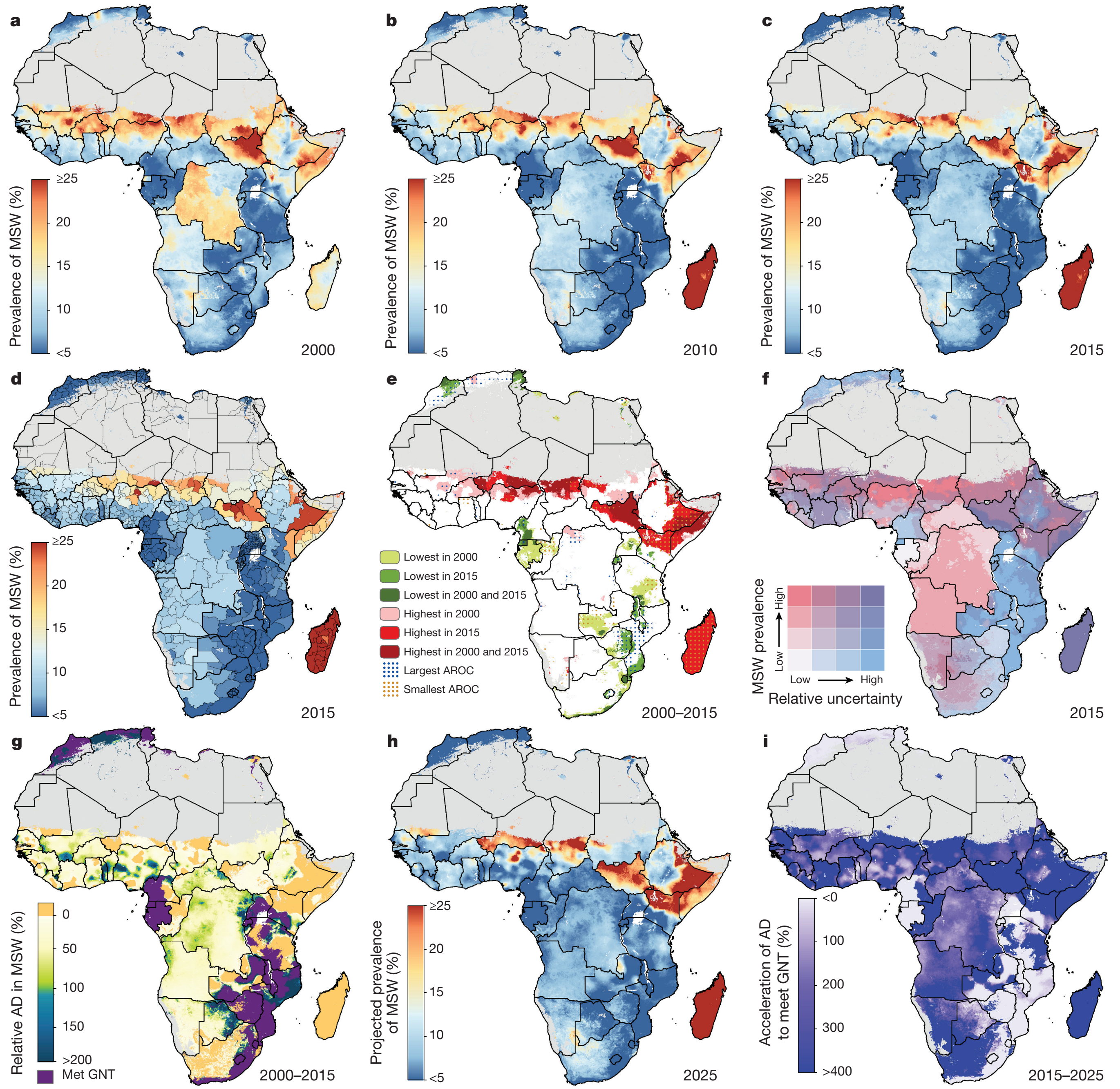


Figure 2 | Wasting prevalence (2000–2015) in children under five and progress towards 2025. a–**c**, Prevalence of moderate and severe wasting (MSW) at the 5 × 5-km resolution in 2000 (**a**), 2010 (**b**) and 2015 (**c**). **d**, Prevalence of wasting at the first administrative subdivision in 2015. **e**, Overlapping population-weighted lowest and highest 10% of pixels and AROC in wasting from 2000 to 2015 across the continent. **f**, Overlapping population-weighted quartiles of wasting and relative 95% uncertainty in 2015. **g**, Annualized decrease in wasting prevalence from 2000 to 2015 relative to rates needed during 2015–2025 to meet the WHO GNT. 100% indicates the annualized decrease from 2000 to 2015 equivalent to the pace of progress required during 2015–2025 to meet the WHO GNT by 2025 (wasting less than 5%). Blue pixels exceeded this pace; green to yellow pixels proceeded at a slower rate than required; orange pixels were non-decreasing; and purple pixels were estimated to have met the target by 2015. **h**, Pixel-level prevalence of wasting was predicted for 2025 on the basis of the annualized decrease achieved from 2000 to 2015 and projected from 2015. **i**, Acceleration in annualized decrease required to meet the WHO GNT by 2025. Purple pixels were either non-decreasing or must accelerate their rate of decline by more than 400% over 2000–2015 rates during 2015–2025 to achieve the target; white pixels require no increase. Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].

(for example, state), second administrative subdivision (for example, district), and 5 × 5-km levels are publicly available in the Global Health Data Exchange (http://ghdx.healthdata.org/record/africa-child-growth-failure-geospatial-estimates-2000-2015) and via bespoke data visualization tools (https://vizhub.healthdata.org/lbd/cgf).

Outlook for the 2025 GNT

Progress towards the WHO GNT has been uneven (Figs 1g, 2g and Extended Data Fig. 2g). The AROC between 2000 and 2015 relative to the pace required to meet the WHO GNT by 2025 clearly shows that some areas are on track or exceeding the pace required to achieve the

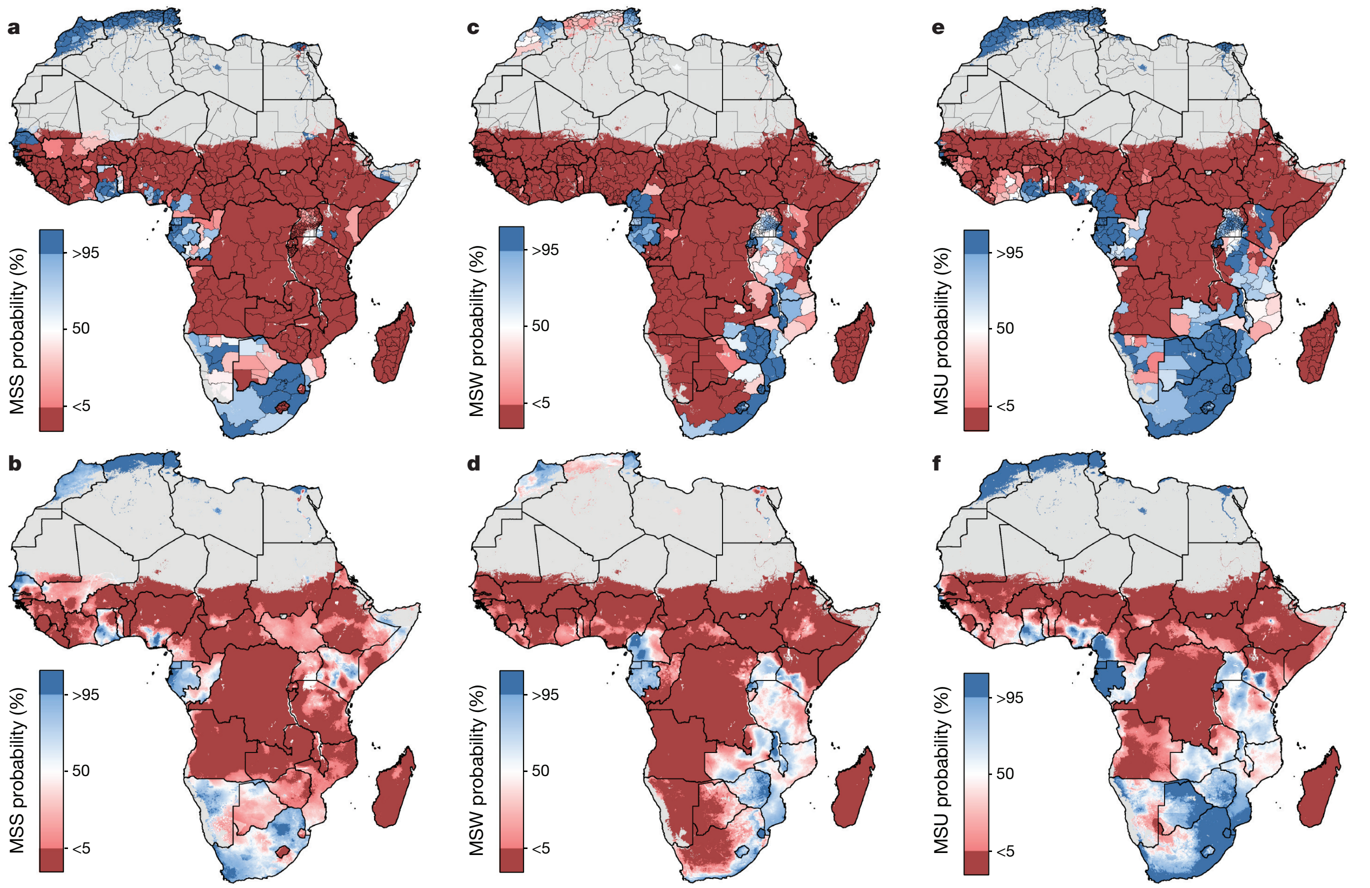


Figure 3 | Probability that the WHO GNT has been achieved in 2015 at the first administrative subdivision and 5 × 5-km pixel level for stunting, wasting and underweight. a–f, Probability of WHO GNT achievement in 2015 at the first administrative subdivision and 5 × 5-km level for moderate and severe stunting (**a**, **b**), moderate and severe wasting (**c**, **d**) and moderate and severe underweight (**e**, **f**). The probability that dark-blue pixels have met the WHO GNT in 2015 is greater than 95%, and less than 5% for dark-red pixels. Estimates for 2015 at the 5 × 5-km level have been calculated using population-weighting based on the population of children under five and probabilities that the WHO GNT were met in 2015. Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].

targets and other locations appear to have already met the goal. Yet there are still vast expanses of the continent that must increase their rate of progress two-, three-, and even fourfold to achieve the WHO GNT by 2025 (Figs 1i, 2i and Extended Data Fig. 2i). Inland areas of central Africa and the Sahel will require the most marked improvement, with many areas requiring at least two- to fourfold increases in their annual rates of decrease across CGF indicators. Stunting targets—40% reduction by 2025—are unlikely to be met in many areas of central Africa without an accelerated rate of decline, while some areas, such as southern and western coastal Africa, are on target to meet the WHO GNT. The reduction and maintenance of child wasting to less than 5% is likely to be met in most countries in southeastern Africa based on current trajectories, but much of central SSA and the entire Sahel will require pronounced improvements in order to meet 2025 targets.

The likelihood of areas meeting the WHO GNT is sensitive to the spatial scale used to measure achievement. In Fig. 3, the probability that regions have reached the WHO GNT in 2015 is estimated at the first administrative subdivision, as well as at the 5 × 5-km pixel level, showing the increased nuance in the results gained at finer spatial scales. For example, the administrative boundaries of Kenya are aligned in a north–south orientation, cutting across areas of low and high prevalence, which generates population-weighted probabilities of having met the stunting target between 0% and 50% for much of eastern Kenya (Fig. 3a). Viewing the same results at a 5 × 5-km resolution (Fig. 3b) shows that pockets within Kenya had a much higher (over 95% in some places) probability of meeting the target in 2015.

Discussion

Our study provides a quantification of CGF in 51 African countries at a 5 × 5 km spatial resolution, highlighting a mixture of impressive gains and enduring disparities in CGF within countries and across the continent. By 2015, nearly all locations showed decreases in the rates of stunting, wasting and underweight compared to 2000, with noticeable tracts of the DRC, Mozambique, Angola and Burkina Faso showing considerable reductions in multiple CGF indicators, despite room for improvement within these locales (Figs 1g, 2g and Extended Data Figs 2g, 3, 4). Conversely, some countries are performing poorly across wide areas in all CGF indicators: South Sudan, Chad, Ethiopia, Madagascar, Sudan and northern Nigeria had some of the highest rates of CGF based on multiple metrics in 2015 (Extended Data Fig. 4). We also provide a baseline assessment of the WHO GNT for CGF at the local level for 2015 (Fig. 3), to guide policy and precision public health interventions to improve outcomes by 2025. Although policies are often set at administrative levels, implementation happens locally, such as within districts or cities, particularly when targeting specific at-risk populations or studying responses to interventions over time. Our analysis at the 5 × 5-km level enables the identification of programme and policy successes and failures at the local level and quantification of inequalities to guide efficient targeting of resources and interventions to those populations with the greatest need.

Differences between alternative sets of international nutrition targets must be acknowledged, along with the fact that even in high-income nations, malnutrition has not been entirely eliminated. MDG 1.C called for a 50% reduction in underweight prevalence between 1990 and 2015. On the basis of national-level GBD results, 21 African countries achieved MDG 1.C and an additional 16 surpassed the achievement rate from 2000 to 2015[1]. The mapping of 5 × 5 km levels and trends in underweight has revealed substantial further geographical heterogeneity during the MDG period (Extended Data Fig. 2h), with almost every country having areas in which improvements were consistent with achieving MDG 1.C, and every country having areas in which improvement lagged behind the national rate. The WHO GNT were formulated by analysing national CGF trends in Brazil, China, Bangladesh and Mexico, all of which made remarkable progress during the MDG period[12]. While most of Africa must accelerate reductions in CGF in children under 5 in order to meet the WHO GNT by 2025, these aspirational goals are anchored in examples of past achievement. By contrast, the current wording of SDG 2.2 calling to 'end all forms of malnutrition' is clinically vague and almost certainly unachievable. There is a need for more clearly defined SDG targets for malnutrition and CGF, formulated in terms of absolute, rather than relative change. Absolute targets would bring SDG 2.2 in line with the overall aim of the SDGs of achieving a 'grand convergence' in health[31].

Most CGF improvements in Africa occurred after the year 2000, and were likely catalysed by large political, social and financial investments[32]. A number of related factors are likely to have led to improvements in nutrition, CGF and child mortality[4]. These include general sociodemographic improvements[9] and broad scaling up of interventions that focused on reducing childhood illness, such as malaria control, vaccination coverage, HIV prevention and treatment, and water, sanitation and hygiene facilities, which can break the cycle of metabolic compromise leading to CGF[9,16,17,33–35]. It is probably no coincidence that many of the nations and regions with slower gains, such as Central African Republic, Chad, Somalia and much of the Sahel, received less international assistance for newborn and child health[26], had persistently low coverage of sentinel maternal and child health interventions[36], experienced periods of pronounced conflict[36,37], and showed no progress in their sociodemographic index status[38]. The finding of a continued high burden of wasting in arid sections of the Sahel, the Horn of Africa and sections of southern SSA is especially important given the implications of famine on potential for human health, geopolitical unrest and mass migration[39,40]. There is strong correspondence between the areas of high prevalence of wasting in 2015 (Fig. 2c) and the nations (Nigeria, South Sudan, Somalia, and Yemen (not mapped here)) that were identified by the United Nations as collectively containing approximately 20 million people that are at imminent risk of famine[41].

We estimate that no country in Africa is likely to achieve all of the WHO GNT in all of its territory if current trends continue, highlighting a widespread need to adopt evidence-based, precision public health programmes to track and improve progress. In an era of static development assistance for health[32] and in countries where financial resources are constrained, highly localized mapping of CGF may facilitate more efficient stewardship by providing a way to pair vulnerable communities with health and nutrition programmes, community support and knowledge that are more likely to meet their specific needs. Targeting precision health interventions to reduce the burden of CGF without considering key sociodemographic factors poses large risks to the sustainability of intervention strategies, either directly through unrealistic assumptions about care-seeking behaviour and retention, or indirectly by not working in tandem to break cycles of poverty and mitigate CGF risk for future generations. Geospatial estimates of average, community-level human capital, such as those provided in the complementary mapping of educational attainment in Africa[42], should be considered when striving to make policy decisions at a local level. The exact combination of intervention packages required for remedial action to combat CGF was not directly addressed in this study. Further context on the diverse range of instruments and interventions to address CGF is provided in the Supplementary Discussion.

Future work and caveats

Our present study offers the analytic framework from which we aim to extend geospatial modelling of CGF to all low- and middle-income countries, with a heightened focus on the modelling of holistic measures of CGF, such as the composite index of anthropometric failure[43]. These more integrated measures would take into consideration the overlapping and longitudinal influences of being born early, born small and having an early childhood characterized by inadequate height or weight gain. To provide a complete baseline and assessment of progress towards all six WHO GNT, we plan to expand our analysis to include mapping of low birth weight, childhood overweight, anaemia in women of reproductive age, and exclusive breastfeeding in the first six months of life.

The accuracy of this work is primarily determined by the volume and fidelity of nationally representative surveys, regardless of the sophistication of the models used. The limitations of these data, including collection biases in anthropometric measurement and non-existent data on deceased children, underscore the need for future refinement and improved data collection. Furthermore, the statistical model does not yet incorporate child-level covariates, which may mask sub-pixel heterogeneity across sex, age and socio-economic factors (see Methods for additional detail on methodological limitations).

National improvements in CGF across Africa may mask large subnational and acute 5 × 5-km grid-level variation, such that no country in SSA has reached the relevant WHO GNT or SDG targets in all of its territory, or is projected to do so by 2025 or 2030, respectively, under current rates of improvement. As researchers, policymakers and programme implementers continue to determine the optimal mix of interventions to alleviate CGF, they now have at their disposal a precision public health tool to monitor subnational inequalities and target interventions to those populations with the greatest need.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 July 2017; accepted 17 January 2018.

1. GBD 2016 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390,** 1345–1422 (2017).
2. Black, R. E. *et al.* Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet* **371,** 243–260 (2008).
3. Pelletier, D. L. & Frongillo, E. A. Changes in child survival are strongly associated with changes in malnutrition in developing countries. *J. Nutr.* **133,** 107–119 (2003).
4. World Health Organization & United Nations Children's Fund. *WHO Child Growth Standards and the Identification of Severe Acute Malnutrition in Infants and Children: A Joint Statement* (WHO Press, 2009).
5. Wang, Y. & Chen, H.-J. in *Handbook of Anthropometry* (ed. Preedy, V. R.) 29–48 (Springer, 2012).
6. World Health Organization. *Comprehensive Implementation Plan on Maternal, Infant and Young Child Nutrition* (WHO Press, 2017).
7. Murray, C. J. L. Shifting to Sustainable Development Goals—implications for global health. *N. Engl. J. Med.* **373,** 1390–1393 (2015).
8. Nilsson, M., Griggs, D. & Visbeck, M. Policy: map the interactions between Sustainable Development Goals. *Nature* **534,** 320–322 (2016).
9. GBD 2016 SDG Collaborators. Measuring progress and projecting attainment on the basis of past trends of the health-related Sustainable Development Goals in 188 countries: an analysis from the Global Burden of Disease Study 2016. *Lancet* **390,** 1423–1459 (2017).
10. GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388,** 1659–1724 (2016).
11. GBD 2013 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386,** 2287–2323 (2015).
12. de Onis, M. *et al.* The World Health Organization's global target for reducing childhood stunting by 2025: rationale and proposed actions. *Matern. Child Nutr.* **9,** 6–26 (2013).

13. International Food Policy Research Institute (IFPRI). *Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030* (IFPRI, 2016).
14. Development Initiatives. *Global Nutrition Report 2017: Nourishing the SDGs.* (Development Initiatives, 2017).
15. Dowell, S. F., Blazes, D. & Desmond-Hellmann, S. Four steps to precision public health. *Nature* **540,** 189–191 (2016).
16. Bhatt, S. *et al.* The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526,** 207–211 (2015).
17. Gething, P. W. *et al.* Mapping *Plasmodium falciparum* mortality in Africa between 1990 and 2015. *N. Engl. J. Med.* **375,** 2435–2445 (2016).
18. Wollum, A., Burstein, R., Fullman, N., Dwyer-Lindgren, L. & Gakidou, E. Benchmarking health system performance across states in Nigeria: a systematic analysis of levels and trends in key maternal and child health interventions and outcomes, 2000–2013. *BMC Med.* **13,** 208 (2015).
19. Roberts, D. A. *et al.* Benchmarking health system performance across regions in Uganda: a systematic analysis of levels and trends in key maternal and child health interventions, 1990–2011. *BMC Med.* **13,** 285 (2015).
20. Khatab, K. Childhood malnutrition in Egypt using geoadditive Gaussian and latent variable models. *Am. J. Trop. Med. Hyg.* **82,** 653–663 (2010).
21. Hagos, S., Hailemariam, D., WoldeHanna, T. & Lindtjørn, B. Spatial heterogeneity and risk factors for stunting among children under age five in Ethiopia: a Bayesian geo-statistical model. *PLoS ONE* **12,** e0170785 (2017).
22. Mtambo, O. P. L., Masangwi, S. J. & Kazembe, L. N. M. Spatial quantile regression using INLA with applications to childhood overweight in Malawi. *Spat. Spatiotemporal Epidemiol.* **13,** 7–14 (2015).
23. Kandala, N.-B., Fahrmeir, L., Klasen, S. & Priebe, J. Geo-additive models of childhood undernutrition in three sub-Saharan African countries. *Popul. Space Place* **15,** 461–473 (2009).
24. ICF. The DHS Program, Data. http://dhsprogram.com/data/ (1998).
25. Soares Magalhães, R. J. S. & Clements, A. C. A. Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria, and helminth infections in West Africa. *PLoS Med.* **8,** e1000438 (2011).
26. Bosco, C. *et al.* Exploring the high-resolution mapping of gender-disaggregated development indicators. *J. R. Soc. Interface* **14,** 20160825 (2017).
27. Kinyoki, D. K. *et al.* Assessing comorbidity and correlates of wasting and stunting among children in Somalia using cross-sectional household surveys: 2007 to 2010. *BMJ Open* **6,** e009854 (2016).
28. Bhatt, S. *et al.* Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalisation. *J. R. Soc. Interface* **14,** 20170520 (2017).
29. UNICEF, WHO & World Bank Group. *Levels and Trends in Child Malnutrition: Joint Child Malnutrition Estimates*; http://www.who.int/nutgrowthdb/estimates2015/en/ (2016).
30. Arlappa, N. & Kokku, S. B. in *Handbook of Public Health in Natural Disasters: Nutrition, Food, Remediation and Preparation* Vol. 10 (eds Watson, R. R. *et al.*) Ch. 20, 335–366 (Wageningen Academic Publishers, 2015).
31. Jamison, D. T. *et al.* Global health 2035: a world converging within a generation. *Lancet* **382,** 1898–1955 (2013).
32. Dieleman, J. L. *et al.* Development assistance for health: past trends, associations, and the future of international financial flows for health. *Lancet* **387,** 2536–2544 (2016).
33. Ng, M. *et al.* Assessing the contribution of malaria vector control and other maternal and child health interventions in reducing all-cause under-five mortality in Zambia, 1990–2010. *Am. J. Trop. Med. Hyg.* **97,** 58–64 (2017).
34. Bhutta, Z. A. *et al.* Countdown to 2015 decade report (2000–10): taking stock of maternal, newborn, and child survival. *Lancet* **375,** 2032–2044 (2010).
35. Jones, G., Steketee, R. W., Black, R. E., Bhutta, Z. A. & Morris, S. S. How many child deaths can we prevent this year? *Lancet* **362,** 65–71 (2003).
36. Leaning, J. & Guha-Sapir, D. Natural disasters, armed conflict, and public health. *N. Engl. J. Med.* **369,** 1836–1842 (2013).
37. Kinyoki, D. K. *et al.* Conflict in Somalia: impact on child undernutrition. *BMJ Glob. Health* **2,** e000262 (2017).
38. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390,** 1151–1210 (2017).
39. Bohle, H. G., Downing, T. E. & Watts, M. J. Climate change and social vulnerability: toward a sociology and geography of food insecurity. *Glob. Environ. Change* **4,** 37–48 (1994). https://doi.org/10.1016/0959-3780(94)90020-5
40. Roberts, L. Nigeria's invisible crisis. *Science* **356,** 18–23 (2017).
41. UN Secretary-General. *Full Transcript of Secretary-General's Joint Press Conference on Humanitarian Crises in Nigeria, Somalia, South Sudan and Yemen*; https://www.un.org/sg/en/content/sg/press-encounter/2017-02-22/full-transcript-secretary-generals-joint-press-conference (2017).
42. Graetz, N. *et al.* Mapping local variation in educational attainment across Africa. *Nature* https://doi.org/10.1038/nature25761 (2018).
43. Svedberg, P. How many people are malnourished? *Annu. Rev. Nutr.* **31,** 263–283 (2011).
44. GeoNetwork. *Global Administrative Unit Layers (GAUL)*; http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691 (2015).
45. LP DAAC. Combined MODIS 5.1 dataset; available at: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12q1 (accessed 1 June 2017).
46. World Wildlife Fund. *Global Lakes and Wetlands Database Level 3* (2004); https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (accessed 1 June 2017)
47. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol. (Amst.)* **296,** 1–22 (2004).
48. World Pop. WorldPop dataset; available at: http://www.worldpop.org.uk/data/get_data/ (accessed 7 July 2017).
49. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4,** 170004 (2017). https://doi.org/10.1038/sdata.2017.4

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was primarily supported by grants OPP1132415 and OPP1093011 from the Bill & Melinda Gates Foundation, as well as grants OPP1106023 to H.G. and OPP1108547 to M.H. M.H. was also supported by funding from Belmont Forum/FACCE-JPI DEVIL Project (NE/M021327/1).

Author Contributions S.I.H. and N.J.K. conceived and planned the study. B.V.P., N.J.H., K.E.W., I.D.L., D.K.K., A.I.M., A.O.-Z., R.W.S., P.Y.L., A.M.N., A.J.B., J.L. and B.K.M. obtained, extracted, processed and geopositioned the data. L.E., D.J.W., H.S.G., N.G. and M.H. constructed covariate data layers. A.O.-Z., N.G., A.I.M. and R.W.S. wrote the computer code and designed and carried out the statistical analyses with input from R.B., R.C.R., J.F.M., D.C.C. and A.D. L.E., A.O.-Z., A.I.M., N.G., R.W.S. and B.V.P. prepared tables and figures. S.I.H., N.J.K., A.O.-Z., C.S., R.W.S., B.V.P., A.I.M., C.J.L.M., S.S.L., A.H.M., A.M.N., L.K.K., E.G.P., R.R., D.M.P., D.L.S., S.E.R. and N.F. provided intellectual inputs into aspects of this study. E.C., P.W.G. and S.B. developed the covariate stacking methodology. A.I.M., A.O.-Z., S.I.H., N.J.K., C.S., R.W.S. and N.G. wrote the first draft of the manuscript, and all authors contributed to subsequent revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.I.H. (sihay@uw.edu) or N.J.K. (nickjk@uw.edu).

Reviewer Information *Nature* thanks H. Chang, E. Giorgi, L. Haddad and B. Reich for their contribution to the peer review of this work.

METHODS

Overview. Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). Our analysis provides estimates of the prevalence of stunting, wasting and underweight in children under 5 (Extended Data Fig. 1) based on univariate growth standards for which age-specific height and weight are benchmarked against children of the same age from healthy reference populations[4,5]. Stunting, wasting and underweight are defined as z scores that are two or more standard deviations below the reference median for height-for-age (HAZ), weight-for-height (WHZ) and weight-for-age (WAZ), respectively. Our primary goal is to provide prevalence predictions across the African continent at a high resolution and we have used methods to provide the best out-of-sample predictive performance at the expense of inferential understanding. We modelled prevalence of each indicator on a 5 × 5-km grid over 51 countries in Africa at an annual resolution from 2000 to 2015. This includes all 48 countries in mainland Africa, as well as islands for which we had survey data, including Madagascar, Comoros, and São Tomé and Príncipe. We do not estimate for island nations for which no available survey data could be sourced, including Mauritius, Seychelles and Cape Verde. After harmonizing the data, we implemented an ensemble modelling framework that feeds into a Bayesian generalized linear model with a correlated space–time error. We took 1,000 draws from the fitted posterior distribution and we combined and processed the draws into 1,000 candidate 5 × 5-km resolution maps that were used to generate all of our final results. The analytical steps and their limitations are described in detail below and additional detail can be found in the Supplementary Information.

Data. We extracted individual-level height, weight and age data for children under 5 from household survey series, including the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Living Standards Measurement Study and Core Welfare Indicators Questionnaire (CWIQ), among other country-specific child health and nutrition surveys[24,50–52]. Each individual record is associated with a cluster, a group of neighbouring households or a 'village' that acts as a primary sampling unit. Some surveys include geographical coordinates or precise place names for each cluster within that survey (50,142 clusters for stunting, 49,564 for wasting and 50,078 for underweight). In the absence of geographical coordinates; coordinates for each cluster, we assigned data to the smallest available administrative areal unit in the survey while correcting for the survey sample design[53,54]. Boundary information for these administrative units was obtained as shapefiles either directly from the surveys or by matching to shapefiles in the Global Administrative Unit Layers[44] database or the Database of Global Administrative Areas[55]. For select cases, shapefiles provided by the survey administrator were used or custom shapefiles were created on the basis of the survey documentation. These areal data were resampled to 10,000 coordinate locations per areal observation using a population-weighted sampling scheme over the relevant area[49]. k-means clustering on the sampled locations reduces the sampled points to a set of k-means centroids acting as proxies for community locations, and the number of points in each cluster informs the weighting given to the point. These centroids are taken to be the geolocations for the observation, and the pseudo-observations are down-weighted in the likelihood evaluation to account for our uncertainty in the precise location of the observation. Weighting by sample size, GPS-located clusters contributed at least 47.4% of the total data per indicator, and resampled areal data contributed the remainder. Extended Data Figures 5, 6 show stunting data availability by type and country from 2000 to 2015. Wasting and underweight data availability can be found in Supplementary Figs 2, 3.

Child anthropometry data. Using the height, weight and age data for each individual, HAZ, WHZ and WAZ were calculated using the age-, sex- and indicator-specific LMS values from the 2006 WHO Child Growth Standards, which takes into account distributional skew using the lambda parameter, the centre of the distribution using the mu parameter, and the spread of the distribution using the sigma parameter[4,5]. The LMS methodology allows for Gaussian z score calculations and comparisons to be applied to skewed, non-Gaussian distributions[56]. These microdata were then collapsed to cluster-level or areal-level prevalence of moderate stunting, wasting and underweight (HAZ < -2s.d., WHZ < -2s.d. and WAZ < -2s.d. below the reference median, respectively). Data from the Somalia Food Security and Nutrition Analysis Unit were provided as already collapsed to cluster-level prevalences (using the WHO 2006 standards).

Data exclusion criteria. Select data sources were excluded for the following reasons: missing survey weights for areal data, missing gender variables, insufficient age granularity (in months) for HAZ and WAZ calculation in children aged 0–2 years, incomplete sampling (for example, only children aged 0–3 years measured), or untrustworthy data (as determined by the survey administrator or by user inspection). Within each source, polygon survey clusters with a sample size of one were excluded. Untrustworthy data refer specifically to the exclusion of six surveys for the reasons described here. Two datasets, the 2009–2010 Ghana Socioeconomic Panel Survey and the 2005 Burkina Faso CWIQ, were excluded because the national prevalence values reported for one or more indicators were determined to be implausibly high based on the country-level trend seen in the seven other Ghana and six other Burkina Faso sources. In addition, the data were only resolved to the first administrative subdivision. This combined with the very coarse spatial resolution makes the data of minor use for our geospatial purposes. Two additional sources, the 2014 MICS Kenya Kakamega and Bungoma surveys, were excluded because, according to the survey documentation, the 'anthropometric data suffered from digit preference for both weight and height', meaning the measurements were rounded with preference for certain numbers in a way that introduced considerable bias. The 2015 Ethiopia Living Standards Measurement Study–Integrated Surveys on Agriculture was excluded, because the low prevalence of child growth failure in the Ogaden region was determined to be unrealistic by specialists in the field of child nutrition. Lastly, the 2015 Egypt Special DHS was excluded because of the non-proportional sample allocation that was designed to estimate the prevalence of hepatitis and certain other non-communicable disease risk factors such that the survey sampling was not equivalent to the rest of the surveys.

Temporal resolution. We estimated prevalence of stunting, wasting and underweight annually from 2000 to 2015 using a model that allows us to account for data points that were continuously measured over time. As such, the model would also allow us to predict at monthly or finer temporal resolutions. However, we are computationally limited by the temporal resolution of our space–time covariates. In order to account for seasonality within each year of observations, periodic splines were fitted to the data by regions defined by GBD[57] (Extended Data Fig. 2).

Seasonality adjustment. Owing to the acute nature of wasting and its relative temporal transience, wasting data were pre-processed to account for seasonality within each year of observation. Generalized additive models (GAMs) were fitted to wasting data across time using the month of interview and a country-level fixed effect as the explanatory variables and WHZ as the response. A 12-month periodic spline for the interview month was used, as well as a spline that smoothed across the whole duration of the dataset and country-level random effects. The GAMs were fitted to the data by regions defined by GBD[57] (Extended Data Fig. 7) in order to allow for different seasonality adjustments across the continent[57]. Once the models were fitted, individual WHZ observations were adjusted, using only the fit from the periodic spline, so that each measurement was consistent with a day that represented a mean day in the periodic spline. The seasonality adjustment introduced relatively little change to the raw data. This analysis could not be run on sources missing interview dates, which were excluded from the wasting data. See Supplementary Information for more detail and the adjustment is shown in Supplementary Figs 5, 6.

Spatial covariates. In order to leverage strength from locations with observations to the entire spatiotemporal domain, we compiled several 5 × 5-km raster layers of possible socio-economic and environmental correlates of CGF in Africa (see Supplementary Table 3 and Supplementary Fig. 4). These covariates were selected on the basis of their potential to be predictive for the set of CGF indicators, after reviewing literature on evidence and plausible hypotheses as to their influence. Acquisition of temporally dynamic datasets, where possible, was prioritized in order to best match our observations and thus predict the changing dynamics of the CGF indicators. Of the 37 covariates included, 23 were temporally dynamic and were reformatted as a synoptic mean over each estimation period or as a mid-period year estimate. The remaining 14 covariate layers were static, and were applied uniformly across all modelling years. Furthermore, we also used a number of covariates that are constant within each country and year: the percentage of population with access to improved toilet types, and per capita lag distributed income, as indicated as predictive of CGF in GBD 2016[1]. Country-level age-standardized mortality rates due to famine as produced by GBD 2016 were also included in the model for wasting. More information, including plots of all covariates, can be found in the Supplementary Information.

An ensemble covariate modelling method was implemented in order to both select covariates and capture possible nonlinear effects and complex interactions between them[28]. For each region, three sub-models were fitted to our dataset, using all of our covariate data as explanatory predictors: GAMs, boosted regression trees and lasso regression. Each sub-model was fitted using fivefold cross-validation to avoid overfitting, and the out-of-sample predictions from across the five holdouts were compiled into a single comprehensive set of predictions from that model. Additionally, the same sub-models were also run using 100% of the data and a full set of in-sample predictions were created. The five sets of out-of-sample sub-model predictions were fed into the full geostatistical model as the explanatory covariates when performing the model fit. The in-sample predictions from the sub-models are used as the covariates when generating predictions using the fitted full geostatistical model. A recent study has shown that this ensemble approach can improve predictive validity by up to 25% over an individual model[28]. More details on the ensemble covariate modelling can be found in the Supplementary Methods and example predictive rasters can be found in Supplementary Fig. 11.

Analysis. *Geostatistical model.* Binomial count data are modelled within a Bayesian hierarchical modelling framework using a logit link function and a spatially and temporally explicit hierarchical generalized linear regression model to fit prevalence of each of our indicators in five regions of Africa as defined in GBD[57] ('Northern', 'Western', 'Southern', 'Central', and 'Eastern'; see Extended Data Fig. 7). The GBD study design sought to create regions on the basis of two primary criteria: epidemiological homogeneity and geographic contiguity[57] (see Extended Data Fig. 7). For each GBD region, we explicitly write the hierarchy that defines our Bayesian model as follows:

$$C_i | p_i, N_i \sim \text{binomial}(p_i, N_i)$$

$$\text{logit}(p_i) = \beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} + \epsilon_{\text{GP}i} + \epsilon_i$$

$$\sum \boldsymbol{\beta} = 1$$

$$\epsilon_i \sim N(0, \sigma^2_{\text{nug}})$$

$$\boldsymbol{\epsilon}_{\text{GP}} | \Sigma_{\text{space}}, \Sigma_{\text{time}} \sim \text{GP}(0, \ \Sigma_{\text{space}} \otimes \Sigma_{\text{time}})$$

$$\Sigma_{\text{space}} = \frac{2^{1-\nu}}{\tau \times \Gamma(\nu)} \times (\kappa \boldsymbol{D})^{\nu} \times K_{\nu}(\kappa \boldsymbol{D})$$

$$\Sigma_{\text{time}_{j,k}} = \rho^{|t_k - t_j|}$$

For each indicator and region, we modelled the number of children at cluster i, among a sample size, $\boldsymbol{N_i}$, who are subject to the indicator as binomial count data, C_i. We have suppressed the notation, but the counts (C_i), probabilities (p_i), predictions from the three submodels (X_i) and residual terms ϵ_* are all indexed at a space–time coordinate. The probabilities (p_i) represent both the annual prevalence at the space–time location and the probability that an individual child will be afflicted with the risk factor given that they live at that particular location. The logit of annual prevalence (p_i) of our indicators was modelled as a linear combination of the three sub-models (GAM, boosted regression trees and lasso regression), $\boldsymbol{X_i}$, a correlated spatiotemporal error term ($\epsilon_{\text{GP}i}$) and an independent nugget effect, ϵ_i. Coefficients (β) on the sub-models represent their respective predictive weighting in the mean logit link and are constrained to sum to 1. In order for this constraint to make any sense, we ensure that the predictions from the sub-models entered into INLA (integrated nested Laplace approximation)[28] in the link space (logit) without having been centre-scaled. The joint error term ($\boldsymbol{\epsilon}_{\textbf{GP}}$) accounts for residual spatiotemporal autocorrelation between individual data points that remains after accounting for the predictive effect of the sub-model covariates, and the nugget (ϵ_i), which is an independent error term for each data point, representing irreducible error for that observation. The residuals ($\boldsymbol{\epsilon}_{\textbf{GP}}$) are modelled as a three-dimensional Gaussian process in space–time centred at zero and with a covariance matrix constructed from a Kronecker product of spatial and temporal covariance kernels. The spatial covariance (Σ_{space}) is modelled using an isotropic and stationary Matérn function[58], and temporal covariance (Σ_{time}) as an annual autoregressive-order-1 function over the 16 years that are represented in the model. This approach leveraged the residual correlation structure of the data to more accurately predict prevalence estimates for locations with no data, while also propagating the dependence in the data through to uncertainty estimates[59]. The posterior distributions were fitted using computationally efficient and accurate approximations in R INLA[60,61] with the stochastic partial differential equations[62] approximation to the Gaussian process residuals. Pixel-level uncertainty intervals were generated from 1,000 draws (that is, statistically plausible candidate maps)[63] created from the posterior-estimated distributions of modelled parameters. Additional detail on the geostatistical model and estimation process can be found in the Supplementary Methods.

To transform pixel-level estimates into a range of information useful for a wide community of potential users, these estimates were aggregated from the 1,000 candidate maps up to the second administrative subdivision, the first administrative subdivision and national levels using population weighted conditional simulation[64]. This aggregation also enabled calibration of estimates to national GBD 2016[1] estimates for 2000, 2005, 2010 and 2015. More details on the calibration can be found in the 'Post estimation' section.

Although the model can predict all locations covered by available raster covariates, all final model outputs for which land cover was classified as 'barren or sparsely vegetated' were masked on the basis of the most recently available MODIS satellite data (2013), as well as areas where the total population density was less than ten individuals per 1 × 1-km pixel in 2015. This step has led to improved understanding of the maps when communicating with data specialists and policymakers.

Post estimation. To leverage national-level data included in GBD 2016, but outside the scope of our current geospatial modelling framework, and to ensure perfect calibration between these estimates and GBD 2016 national-level estimates, we performed a post hoc calibration to each of our 1,000 candidate maps[1]. For each posterior draw, we calculated population-weighted pixel aggregations to a national level and compared these country–year estimates to the analogous and available GBD 2016[1] country–year estimates (all countries for 2000, 2005, 2010 and 2016). To generate 2015 national-level estimates for use in calibrating our 2015 5 × 5-km maps, we linearly interpolated between 2010 and 2016 estimates. We defined the raking factor to be the ratio between the GBD 2016[1] estimate and our current estimates and linearly interpolated raking factors in a country between the available years yielding raking factors for all country–year pairs. Finally, we multiplied each of our pixels in a country–year pair by its associated raking factor. This ensures perfect calibration between our geospatial estimates and GBD 2016[1] national-level estimates, while preserving our estimated within-country geospatial and temporal variation.

The median for the raking factor ratios across all three indicators was 0.999 (interquartile range, 0.920–1.096), indicating a very close agreement with GBD 2016[1] estimates. Scatter plots comparing national-level estimates from this analysis with GBD 2016[1] estimates can be found in Supplementary Figs 40–42.

Model validation. Models were validated using spatially stratified fivefold out-of-sample cross-validation. In order to offer a more stringent analysis by respecting some of the spatial correlation in the data, holdout sets were created by combining sets of spatially contiguous data at different spatial resolutions, for example, the first administrative subdivision. Validation was performed by calculating bias (mean error), total variance (root-mean-square error) and 95% data coverage within prediction intervals, and correlation between observed data and predictions. All validation metrics were calculated on the out-of-sample predictions from the fivefold cross-validation. We compared five different model formulations (stacked ensemble with and without space–time error, raw satellite covariates with and without space–time error, and the Gaussian process space–time error without any covariates) using out-of-sample predictive metrics. The results are presented in the model validation section of the Supplementary Methods, in which we show that using the stacked ensemble covariates in conjunction with the space–time error consistently outperforms the other models across all three indicators.

Where possible, results from these models were compared against other existing estimates, such as subnational DHS estimates as shown in Supplementary Fig. 43. Furthermore, measures of spatial and temporal autocorrelation pre- and post-modelling were examined to verify correct recognition, fitting and accounting for the complex spatiotemporal correlation structure in the data. We found our in-sample-size weighted Pearson's correlation between our posterior mean predictions at data observation locations and the observed prevalence proportions to be 0.70, 0.66 and 0.76 for stunting, wasting and underweight, respectively, at the pixel level, and 0.98, 0.96 and 0.99, respectively, at the national level. The equivalent out-of-sample correlations were 0.63, 0.58 and 0.69 for stunting, wasting and underweight, respectively, at the pixel level, and 0.96, 0.95 and 0.98, respectively, at the national level. We also used various out-of-sample validation strategies to assess the fit of our models. For example, for stunting we demonstrate that our models, aggregated to the national level over five-year periods, have a small average root mean square error (0.020, ranging from 0.017 to 0.023), a small average mean error (0.0175, 0.001–0.012), a well-calibrated average 95% coverage (93.25%, ranging from 91.6% to 94.3%) and a high concordance with existing small area estimates (Supplementary Fig. 31). All model validation procedures and corresponding results are provided in the Supplementary Methods.

Projections. To compare our estimated rates of improvement in CGF prevalence over the last 15 years with the improvements needed between 2015 and 2025 to meet the WHO GNT, we performed a simple projection using estimated AROCs applied to the final year of our estimates. A full predictive forecast was not available due to a lack of available forecasts for many of our covariates.

For each CGF indicator i, we calculated log-additive annual rates of change at each pixel j, by logit-transforming our 16 years of posterior mean prevalence estimates, $\text{prev}^l_{i,j,\text{yr}}$, and calculating the annual rate of change between each pair of adjacent years starting with 2001:

$$\text{AROC}^l_{i,j,\text{yr}} = \text{prev}^l_{i,j,\text{yr}} - \ \text{prev}^l_{i,j,\text{yr}-1}$$

We then calculated a weighted AROC for each indicator–pixel by taking a weighted average across the years, where more recent AROCs are given more weight in the average. We defined the weights to be:

$$W_{\text{yr}} = \frac{(\text{yr} - \ 2000)^{\gamma}}{\sum_{2001}^{2015} (\text{yr} - 2000)^{\gamma}}$$

where γ may be chosen to give varying amounts of weight across the years. For this set of projections we selected $\gamma = 1$, resulting in a linear weighting scheme that has been tested and vetted for use in projecting the health-related SDGs[9]. For any indicator and for any pixel, we then calculated the average AROC to be:

$$\mathrm{AROC}_{i,j} = \sum_{2001}^{2015} W_{\mathrm{yr}} \mathrm{AROC}^{l}_{i,j,\mathrm{yr}}$$

Finally, we calculated the projections by applying the ten years of the annual rates of change at each pixel in our mean 2015 mean prevalence estimates:

$$\mathrm{Proj}_{i,j,2025} = \mathrm{logit}^{-1}(\mathrm{prev}^{l}_{i,j,2015} + \mathrm{AROC}_{i,j} \times 10)$$

This projection scheme is analogous to the methods used in the GBD 2016 measurement of progress and projected attainment of health-related SDGs[9]. An evaluation of the projection methodology and the implicit assumptions involved can be found in the Supplementary Methods.

Relative WHO GNT interpretation. The WHO GNT are composed of both relative (for example, 40% reduction in stunting relative to 2010) and fixed (for example, less than 5% wasting) targets. In order to compare our modelled results to the relative WHO GNT, we computed the population-weighted aggregated prevalence in 2010 from GBD 2016[1] results across all countries for which we made estimates. We then set a fixed target for every pixel in our modelled domain to be a reduction based on the 2010 continent-level aggregated prevalences. This interpretation of the WHO GNT was used to set a fixed target across space while ensuring that locations that were already performing favourably were not characterized as being behind pace to reach the targets due to their early and continued low prevalences across time. This yielded a stunting prevalence target of 24.2%, and an underweight target prevalence of 13.5%.

Limitations. This work should be assessed in full acknowledgement of the data and methodological limitations. While our present study is informed by 209 sources (totalling 1.29 million measured children), areas of greatest uncertainty (Figs 1f, 2f and Extended Data Fig. 2f) usually correspond to those in need of newer and/or updated information (Extended Data Figs 5, 6 and Supplementary Figs 2, 3). Expansion to additional countries and indicators underscores the need for enhanced data collection (and equally importantly, retrospective data retrieval) as we iteratively update the measurement of progress towards global targets. While not a focus of this study, a combination of the magnitude of CGF indicator prevalence (Figs 1c, 2c and Extended Data Fig. 2c), the uncertainty in its estimation (Figs 1f, 2f and Extended Data Fig. 2f), and our knowledge of national survey coverage (Extended Data Figs 5, 6 and Supplementary Figs 2, 3) can be used to help to identify countries and vulnerable sub-populations that would benefit from further survey enumeration.

There are limitations to the data used in this analysis and thus areas for future refinement. For example, the height or weight of children may have been measured or recorded incorrectly due to equipment calibration or user error, or based on difficulties originating from measuring younger children lying down rather than standing up[65]. Levels of 'missingness' in these survey data may also be high due to recall error of a child's birthday. Given that growth standards are age- and sex-specific, children without detailed age information were excluded from the analysis (see Supplementary Information). In addition, a child must have been present in the home in order for the survey taker to record measurements. Given that only children alive at the time of the survey could be counted, children under 5 who died due to undernourishment or other causes before the survey was taken would not have been measured. Conflict zones in select countries or regions may also have been excluded from surveying because of security and safety issues. The direction of all of these biases is towards an underestimation of CGF.

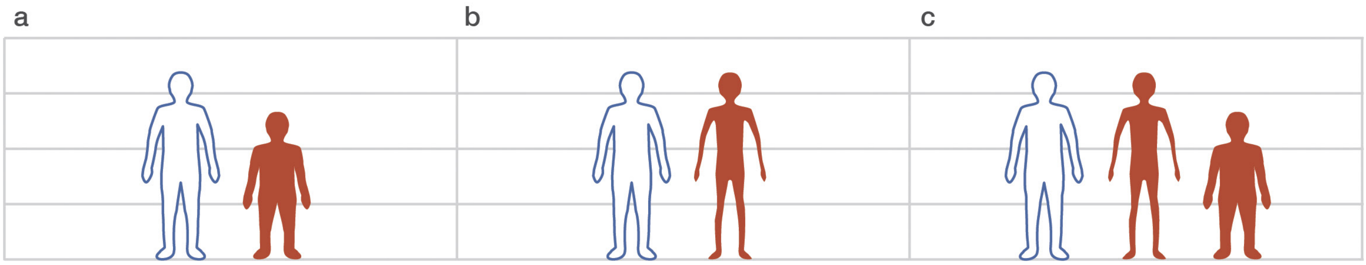
Moreover, our estimates are not stratified by sex, wealth or any other socio-economic indicators. This may mask higher rates of CGF present in sub-populations within the areas measured and while this work presents a very fine scale for comprehensive geospatial estimates of child growth failure, the 5 × 5-km resolution is still too coarse to account for urban slums and other hyperspecific spatial disparities. Similarly, relatively coarse AROCs taken across time may obscure higher-frequency changes within the time series, and more research in studying and summarizing spatially correlated temporal trends should be pursued. Although comprehensive, due to a lack of high-resolution spatial data, our set of included covariates does not cover all CGF drivers and confounders. On the modelling side, we have attempted to propagate as much uncertainty through the various modelling stages, but there are still some propagations, such as incorporating uncertainty from the child model ensemble fits, that proved computationally infeasible. Future research is also ongoing to develop computational methods for better geostatistical integration of point and areal data to continental-scale mapping studies for a variety of indicators. These geostatistical tools are driven primarily by infrequently reported national survey data and are thus well-positioned for monitoring and evaluating progress across years, but are not suited for day-to-day assessments of CGF vulnerability. We show, however, that there is considerable room for exciting harmonization with such efforts, for example, by focusing attention of early warning efforts on populations that are the most vulnerable and least resilient[66].

Code availability. All code used for these analyses is publicly available online at http://ghdx.healthdata.org/record/africa-child-growth-failure-geospatial-estimates-2000-2015.

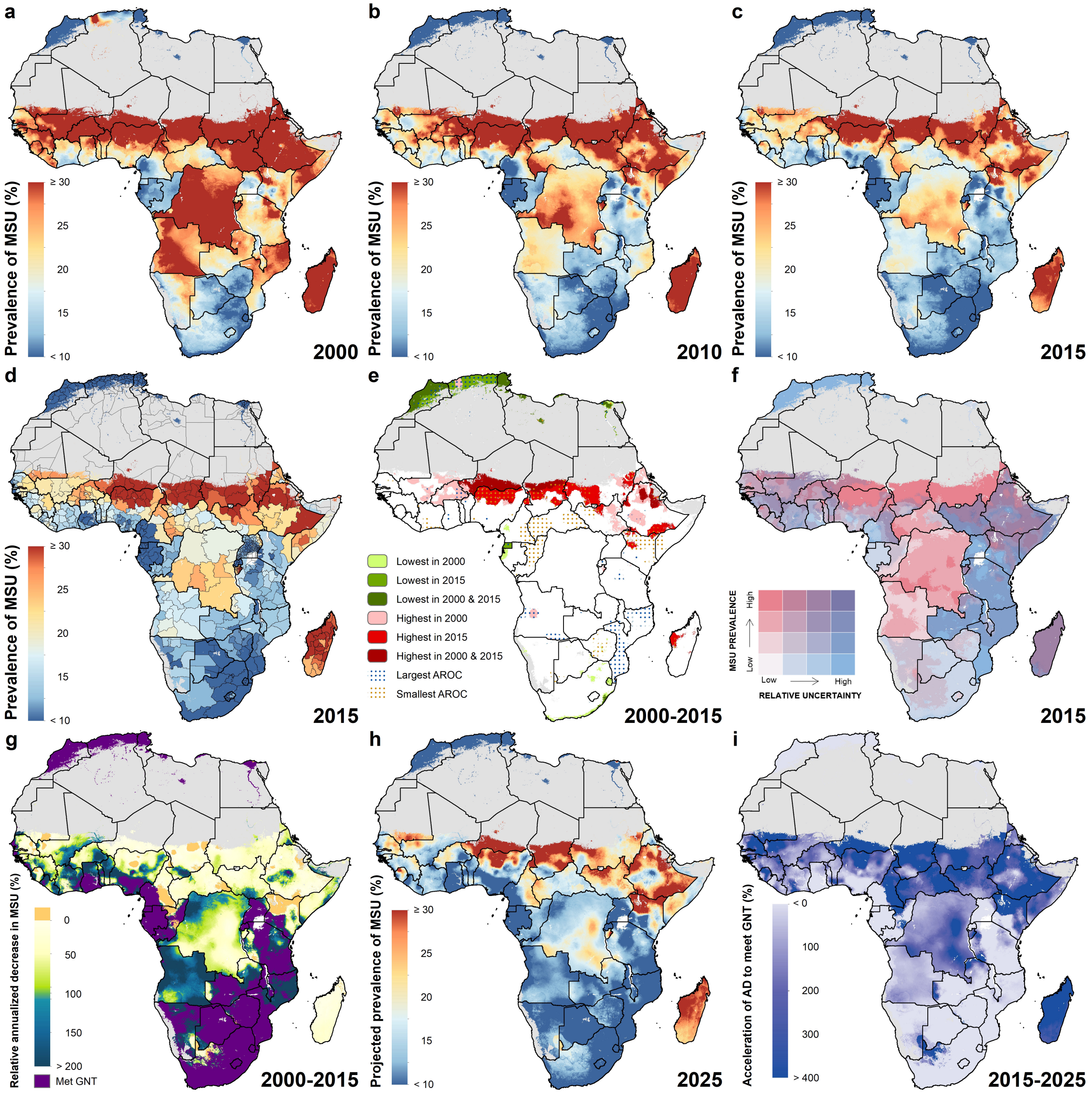
Data availability. The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider, and data that are not publicly available due to restrictions by the data provider, which were used under license for the current study, but may be available from the authors upon reasonable request and permission of the data provider. A detailed table of data sources and availability can be found in Supplementary Table 2.

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by the FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects[44]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center[45]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel[46,47]. Populations were retrieved from WorldPop[48,49].

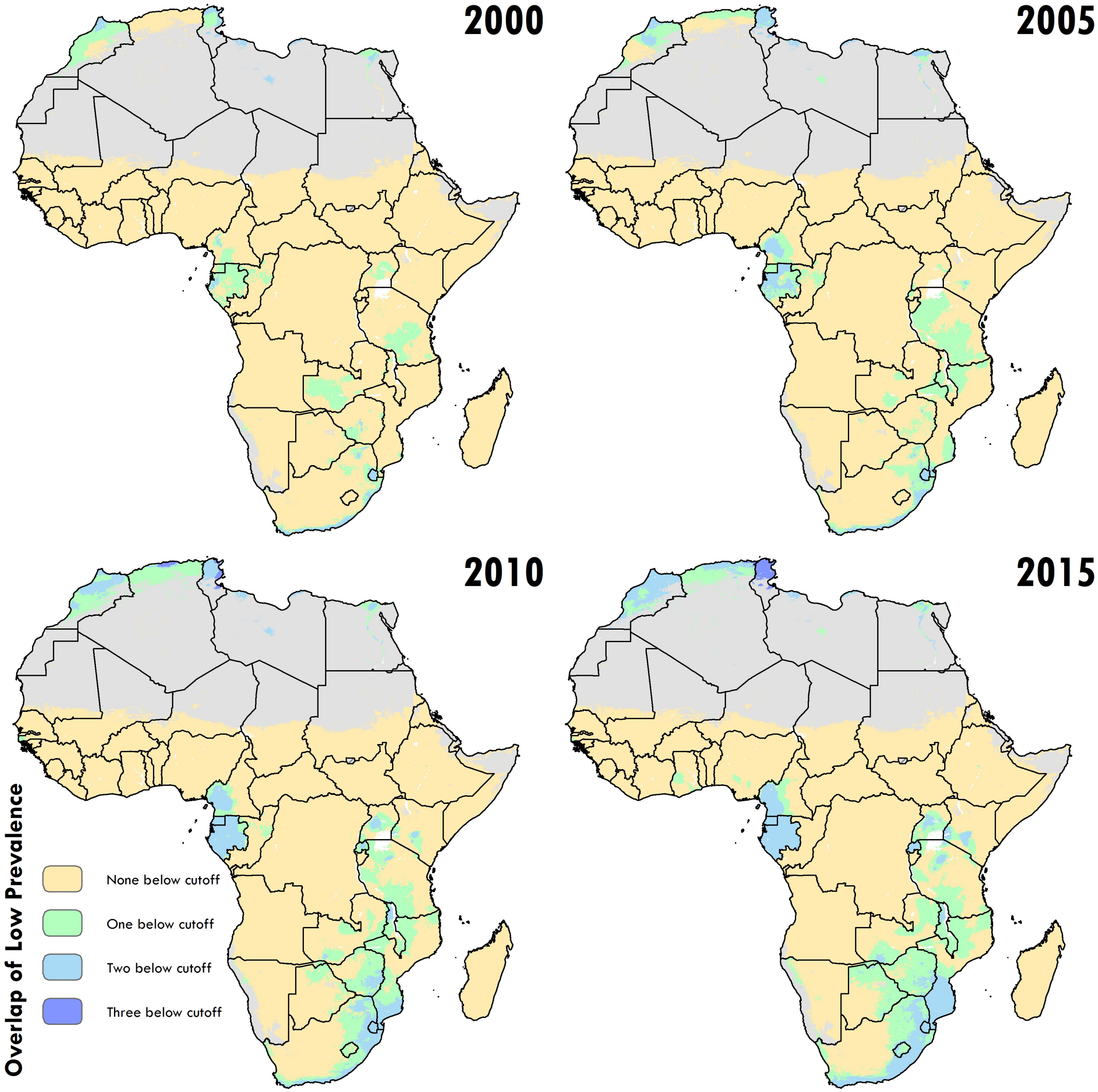
50. UNICEF. *Multiple Indicator Cluster Survey (MICS);* https://www.unicef.org/statistics/index_24302.html (UNICEF, 2010).
51. World Bank Group. *Living Standards Measurement Survey (LSMS);* http://go.worldbank.org/UK1ETMHBN0 (World Bank Group, 2016).
52. Global Health Data Exchange. *Core Welfare Indicators Questionnaire Survey (CWIQ);* http://ghdx.healthdata.org/series/core-welfare-indicators-questionnaire-survey-cwiq (World Bank, accessed 21 April 2017).
53. Lumley, T. in *Complex Surveys* (eds Couper, M. P. *et al.*) 17–37 (John Wiley & Sons, 2010).
54. Lumley, T. Analysis of complex survey samples. *J. Stat. Softw.* **9,** https://doi.org/10.18637/jss.v009.i08 (2004).
55. Global Administrative Areas. *GADM Database Of Global Administrative Areas* version 2.8 http://www.gadm.org/ (2015).
56. Indrayan, A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr.* **51,** 37–43 (2014).
57. Murray, C. J. *et al.* GBD 2010: design, definitions, and metrics. *Lancet* **380,** 2063–2066 (2012).
58. Stein, M. L. *Interpolation of Spatial Data* (Springer New York, 1999).
59. Waller, L. & Carlin, B. in *Handbook of Spatial Statistics* (eds Gelfand, A. *et al.*) 217–243 (CRC, 2010).
60. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71,** 319–392 (2009).
61. Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* **67,** 68–83 (2013).
62. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. B* **73,** 423–498 (2011).
63. Patil, A. P., Gething, P. W., Piel, F. B. & Hay, S. I. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol.* **27,** 246–253 (2011).
64. Gething, P. W., Patil, A. P. & Hay, S. I. Quantifying aggregated uncertainty in *Plasmodium falciparum* malaria prevalence and populations at risk via efficient space-time geostatistical joint simulation. *PLOS Comput. Biol.* **6,** e1000724 (2010).
65. Assaf, S., Kothari, M. T. & Pullum, T. *An Assessment of the Quality of DHS Anthropometric Data, 2005–2014.* DHS Methodological Report 16 (ICF International, 2015).
66. FEWS NET. *Famine Early Warning Systems Network*; https://www.fews.net/ (accessed 28 April 2017).



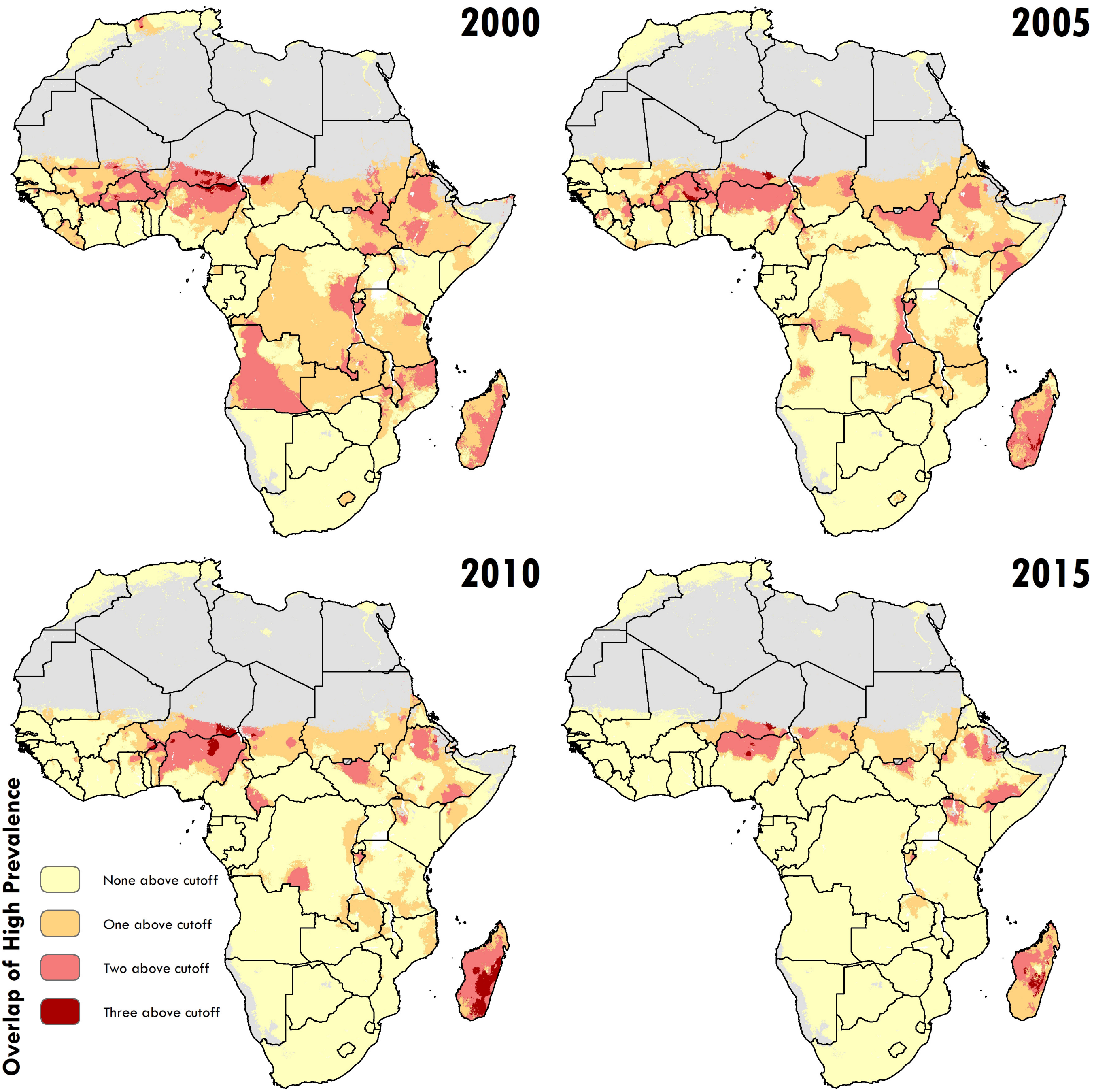
Extended Data Figure 1 | Measurement of child growth failure. **a**, Stunting is a manifestation of chronic malnutrition and is defined as a height-for-age *z* score (HAZ) that is two or more standard deviations (s.d.) below the reference median. **b**, Wasting is an emaciated state resulting from acute malnutrition and is defined a weight-for-height *z* score (WHZ) of <-2. **c**, Underweight is a weight-for-age *z* score (WAZ) of <-2 and is considered a marker of subacute malnutrition, but is nonspecific from an anthropometric standpoint, because it can indicate either low weight for height, low height for age or some combination of both. There are multiple permutations of child growth failure and the silhouettes are simply illustrative of what a stunted, wasted or underweight child may look like. The World Health Organization Global Targets 2025 to improve maternal, infant and young child nutrition call for a 40% reduction in stunting and a reduction and maintenance of child wasting to less than 5% in children under five. While there is no target for child underweight, a reduction of 40% was used in this analysis.



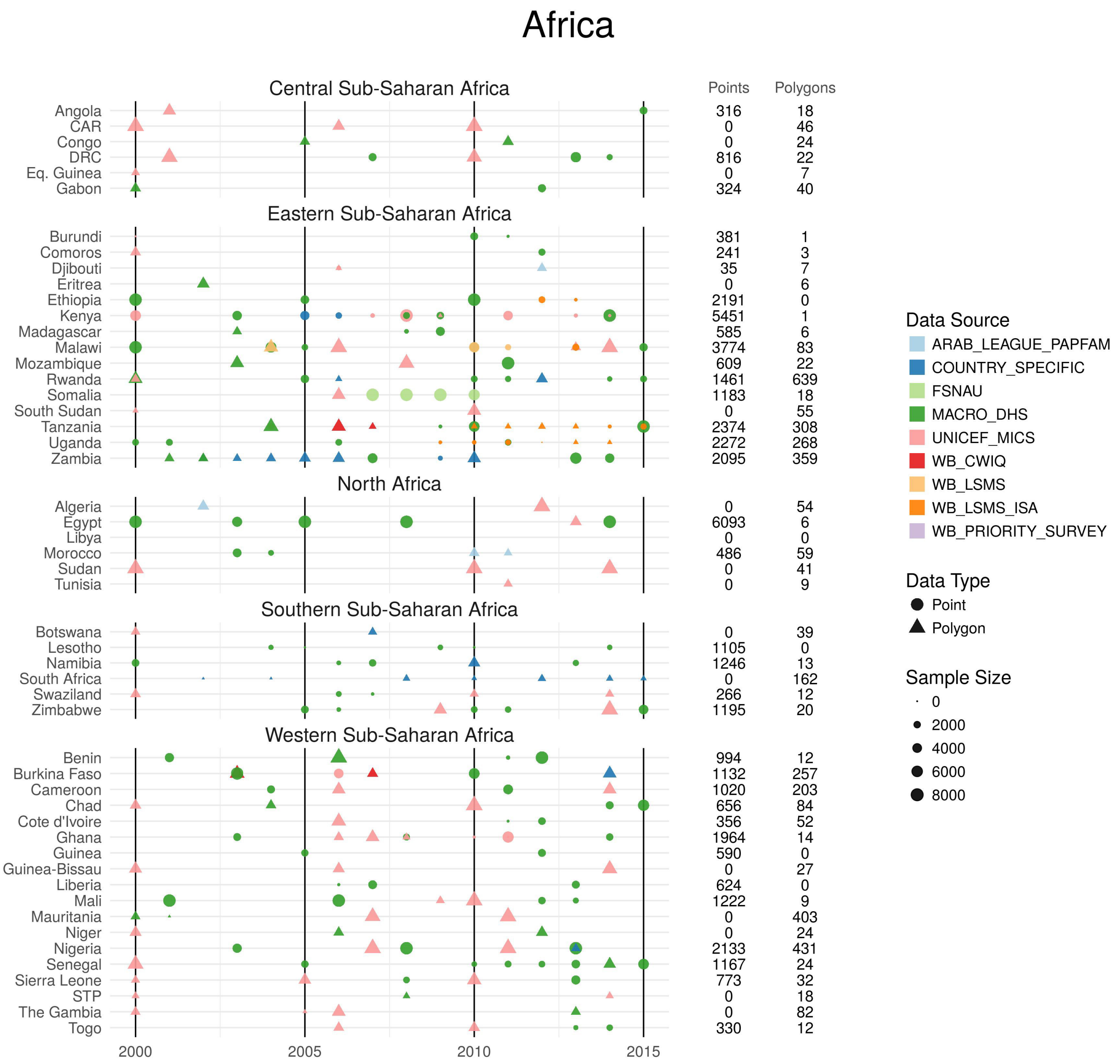
Extended Data Figure 2 | Underweight prevalence in children under five (2000–2015) and progress towards 2025. a–c, Moderate and severe underweight (MSU) prevalence at the 5 × 5-km resolution in 2000 (**a**), 2010 (**b**) and 2015 (**c**). **d**, Underweight prevalence at the first administrative subdivision in 2015. **e**, Overlapping population-weighted lowest and highest 10% of pixels and annualized rates of change in underweight from 2000 to 2015 across the continent. **f**, Overlapping population-weighted quartiles of underweight and relative 95% uncertainty in 2015. **g**, Annualized decrease in underweight prevalence from 2000 to 2015 relative to rates needed during 2015–2025 to meet the WHO GNT. 100% indicates the annualized decrease from 2000 to 2015 equivalent to the pace of progress required during 2015–2025 to meet a 40% decrease in underweight by 2025, relative to 2010. Blue pixels exceeded this pace; green to yellow pixels proceeded at a slower rate than required; orange pixels were non-decreasing; and purple pixels were estimated to have met the target by 2015. This target was internally constructed, commensurate with the target for stunting, as there is no WHO GNT for underweight. **h**, Pixel-level underweight prevalence was predicted for 2025 on the basis of the annualized decrease achieved from 2000 to 2015 and projected from 2015 estimates. **i**, Acceleration in annualized decrease required to meet the WHO GNT by 2025. Purple pixels were either non-decreasing or must accelerate their rate of decline by more than 400% over 2000–2015 rates during 2015–2025 to achieve the target; white pixels require no increase. Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as ‘barren or sparsely vegetated’ are coloured in grey[44–49].



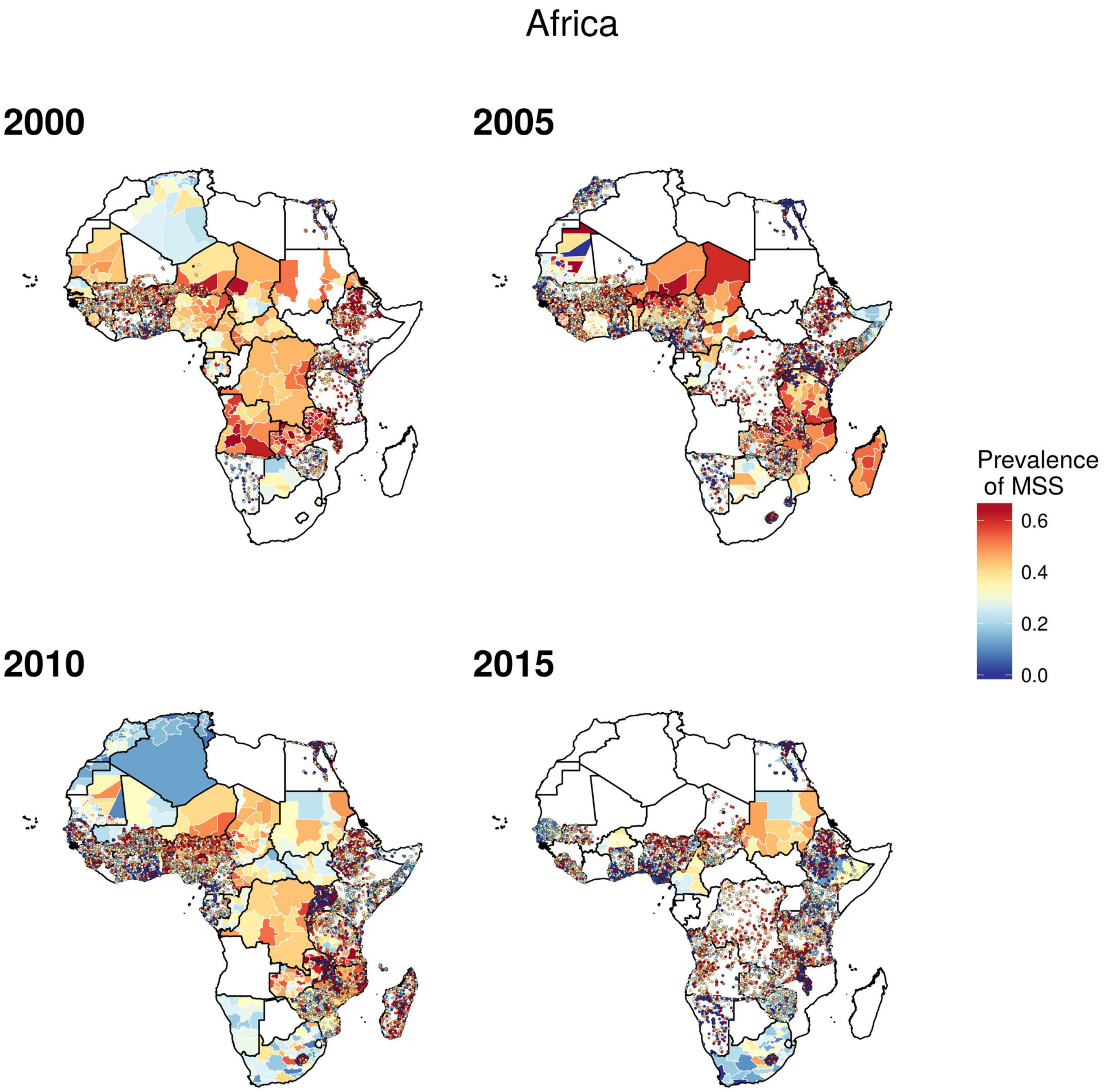
Extended Data Figure 3 | Low prevalence across stunting, wasting and underweight. Across the modelling regions and 5-year periods, these plots show locations where the prevalence of one, two or three of the indicators falls below a lower bound (10% for stunting (HAZ), 5% for wasting (WHZ) and 10% for underweight (WAZ), which correspond to the lower cut-offs used in Figs 1a, 2a, Extended Data Fig. 2a). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].



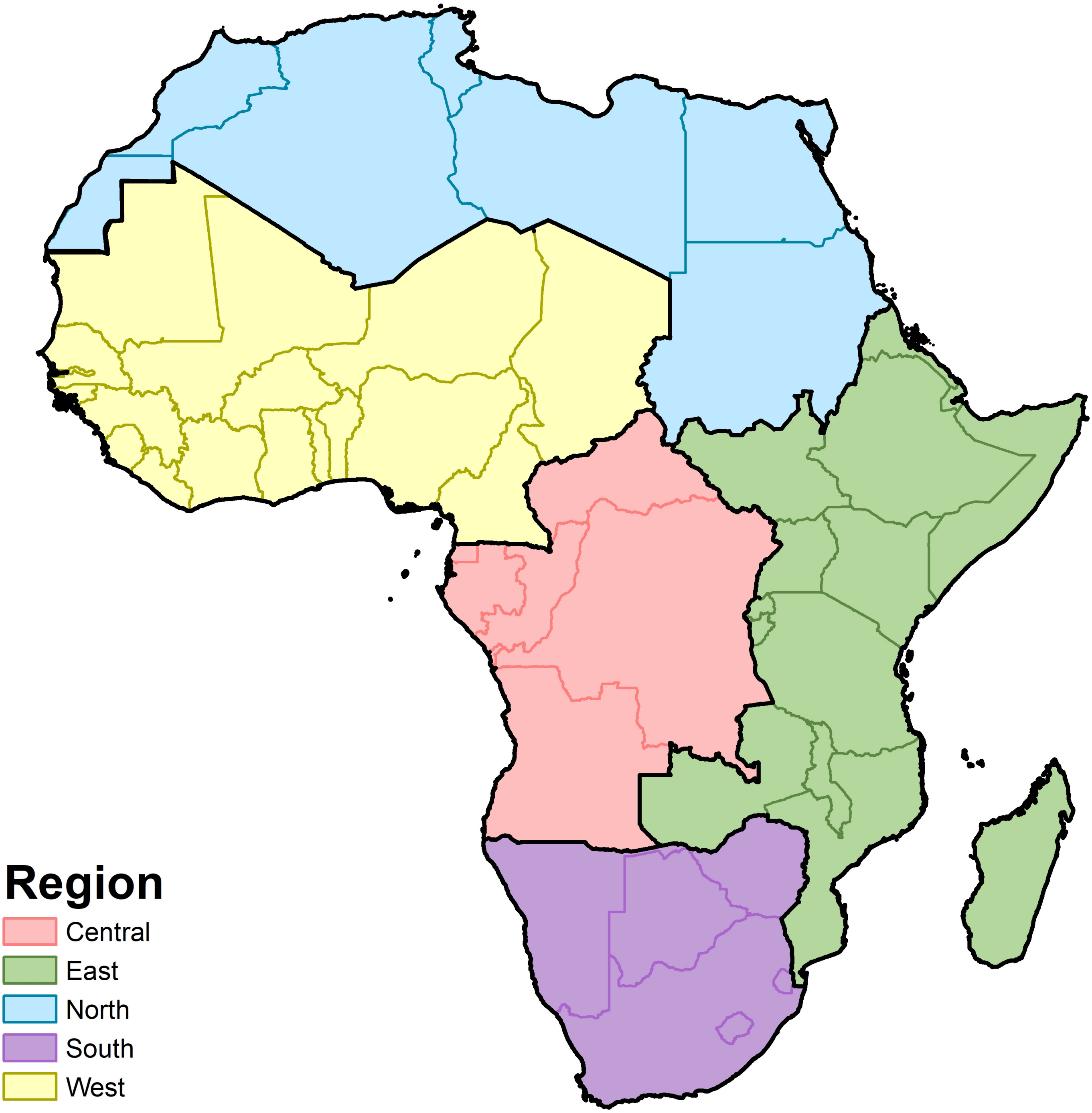
Extended Data Figure 4 | High prevalence across stunting, wasting, and underweight. Across the modelling regions and 5-year periods, these plots show locations where the prevalence of one, two or three of the indicators falls above an upper bound (50% for stunting (HAZ), 25% for wasting (WHZ) and 30% for underweight (WAZ), which correspond to the upper cut-offs used in Figs 1a, 2a, Extended Data Fig. 2a). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].



Extended Data Figure 5 | Stunting annual data availability by type and country for 2000–2015. All data are shown by country and year of survey. The total number of points and polygons (areal) for each country are plotted by data source, type and sample size. Sample size represents the number of individual microdata records for each survey. This database consists of 50,142 clusters and 4,253 polygons with a sample size totalling over 1.15 million children in Africa.



Extended Data Figure 6 | Stunting data availability map for 2000–2015. All data are shown by country and year and mapped at their corresponding geopositioned coordinate or area. Mean stunting prevalence of the input coordinate or area is mapped. This database consists of 50,142 clusters and 4,253 polygons with a sample size totalling over 1.15 million children in Africa. Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[44–49].



Extended Data Figure 7 | Map of GBD regions. Modelling regions were defined as the five GBD regions of Central (central SSA), East (eastern SSA), North (North Africa and the Middle East), South (southern SSA) and West Africa (western SSA)[57]. As this study was limited to mainland Africa and African island nations, select countries were excluded from the North Africa and Middle East region (Afghanistan, Bahrain, Iran, Iraq, Jordan, Kuwait, Lebanon, Oman, Palestinian territories, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, and Yemen). Western Sahara was included as part of the North region.

OPEN
doi:10.1038/nature25761

Mapping local variation in educational attainment across Africa

Nicholas Graetz[1], Joseph Friedman[1], Aaron Osgood-Zimmerman[1], Roy Burstein[1], Molly H. Biehl[1], Chloe Shields[1], Jonathan F. Mosser[1], Daniel C. Casey[1], Aniruddha Deshpande[1], Lucas Earl[1], Robert C. Reiner Jr[1], Sarah E. Ray[1], Nancy Fullman[1], Aubrey J. Levine[1], Rebecca W. Stubbs[1], Benjamin K. Mayala[1], Joshua Longbottom[2], Annie J. Browne[2], Samir Bhatt[3], Daniel J. Weiss[2], Peter W. Gething[2], Ali H. Mokdad[1], Stephen S. Lim[1], Christopher J. L. Murray[1], Emmanuela Gakidou[1]§ & Simon I. Hay[1,2]§

Educational attainment for women of reproductive age is linked to reduced child and maternal mortality, lower fertility and improved reproductive health. Comparable analyses of attainment exist only at the national level, potentially obscuring patterns in subnational inequality. Evidence suggests that wide disparities between urban and rural populations exist, raising questions about where the majority of progress towards the education targets of the Sustainable Development Goals is occurring in African countries. Here we explore within-country inequalities by predicting years of schooling across five by five kilometre grids, generating estimates of average educational attainment by age and sex at subnational levels. Despite marked progress in attainment from 2000 to 2015 across Africa, substantial differences persist between locations and sexes. These differences have widened in many countries, particularly across the Sahel. These high-resolution, comparable estimates improve the ability of decision-makers to plan the precisely targeted interventions that will be necessary to deliver progress during the era of the Sustainable Development Goals.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) states that the ultimate mission of the education targets in Sustainable Development Goal (SDG) 4 is to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all"[1–3]. This is important, because it has been shown that increasing the number of years of schooling that are completed (educational attainment), can lead to higher capital, greater social mobility and increased equity among men and women, in these and other socio-economic outcomes[1,2,4–8]. Educational attainment for women of reproductive age is also among the leading social determinants of health, with higher attainment being strongly associated with improved reproductive health and decreased child mortality[9–14]. The causal pathway between education and health is difficult to study, because randomized control trial methods are logistically challenging and ethically problematic. Observational studies controlling for other predictors of health status, such as age and income, however, indicate that even small gains in educational attainment may improve health outcomes across a wide variety of low-income contexts. Studies across diverse settings have found that increased education for women of reproductive age is associated with improved child nutrition and decreased child mortality, and this effect is consistently stronger than increases in income[15,16]. Importantly, a comprehensive multi-level study found that increases in average attainment in communities are associated with improved survival for infants born to all women in that community, regardless of their own educational attainment or income[17]. This is consistent with research on health behaviours, showing that less-educated women model health behaviours on those of their broader community[18]. These improved health outcomes have also been shown through increased use of prenatal care, greater adherence to treatment regimens and increased contraception use[9,12,19,20]. Despite these clear benefits, international aid for basic education has been deprioritized as a proportion of total aid expenditure every year since 2010[21].

Precision public health and education

SDG 4 focuses on the reduction of inequalities in education on the basis of factors such as wealth, sex and location[1,2,22]. In addition, UNESCO's agenda for reforming education access in developing countries is itself centred around equity[22,23]. Global health efforts have included substantial investments in the use of data to guide interventions that will benefit populations more efficiently and increase equity in outcomes, a strategy that has been termed precision public health[24]. The same paradigm should be extended to the social determinants of health that must be addressed for progress to be sustained. Therefore, although comparable indicators of educational attainment exist at the national level, it is increasingly important to measure subnational variation.

While past studies have assessed subnational variation in attainment for specific African countries[25,26], to our knowledge no comprehensive and comparable set of estimates exist for the continent. Here we build a precisely geolocated database of 173 unique census and survey sources containing information on educational attainment (see Supplementary Figs 1–4 and Supplementary Table 2 for information on data type, coverage and source). We estimate the average number of years of attainment for women of reproductive age (15–49) across a grid of 5 × 5 km across 51 countries in Africa from 2000 to 2015. We also estimate attainment for 20–24-year-old women to more closely identify changes over time. Finally, we construct equivalent models for men to examine differences between the sexes at the same local level. We use recently developed Bayesian spatiotemporal methods[27–29] for the analysis of this dataset, leveraging the high-resolution spatial and temporal information from these data. The estimates produced by these models enable comparisons of subnational regions. We focus on geographical inequality at the 5 × 5-km or local level to explore the subnational distribution of educational attainment, for the following reasons. First, data are increasingly geolocated to specific

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington 98121, USA. [2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7FZ, UK. [3]Department of Infectious Disease Epidemiology, Imperial College London, London SW7 2AZ, UK.
§These authors jointly supervised this work.

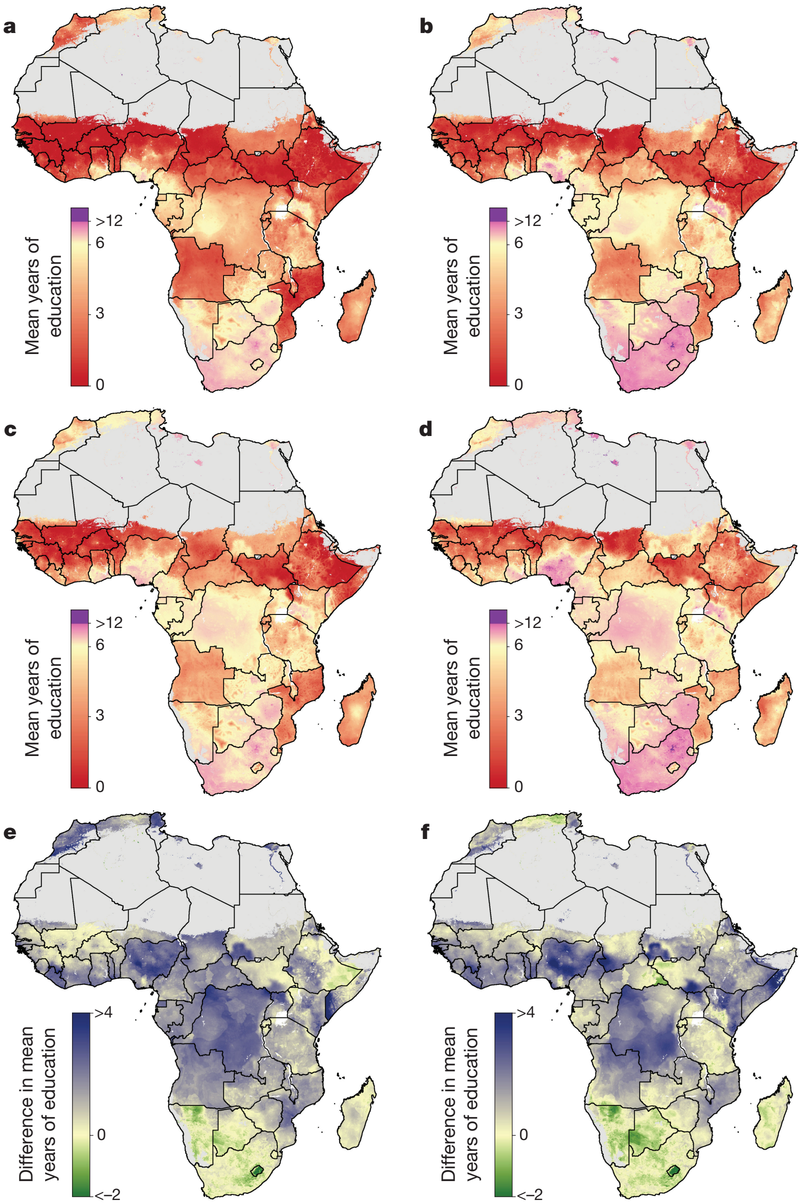


Figure 1 | Average educational attainment for and absolute difference between women and men aged 15–49 in 2000 and 2015. a–d, Average educational attainment for women (**a**, **b**) and men (**c**, **d**) aged 15–49 in 2000 (**a**, **c**) and 2015 (**b**, **d**). **e**, **f**, The absolute difference in average educational attainment between men and women aged 15–49 in 2000 (**e**) and 2015 (**f**). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[32,36–40].

communities, and advances in Bayesian model-based geostatistics enable the modelling of these precise space–time covariance structures. Second, through the increasing availability of satellite imagery and other geospatial modelling endeavours, we have built a collection of covariates at the 5 × 5-km scale that are included in this predictive modelling framework. These are mostly available at only the community level, but allow us to predict outside of our data to estimate mean educational attainment and its uncertainty across all of Africa as a guide for policy formulation and intervention targeting. The utility of community-level and individual-level measurements is discussed in the Supplementary Discussion.

Persistent differences in educational attainment

We used various validation strategies to assess the fit of our models. Across Africa, we use out-of-sample cross-validation to demonstrate that our models have low root mean square errors, low absolute errors, well-calibrated coverages and high concordance with existing small-area estimates (see Supplementary Figs 12–28, Supplementary Tables 8–23). Estimates of mean years of educational attainment for men and women aged 15–49 and 20–24 are shown in Fig. 1a–d and Fig. 2a–d, respectively. These summaries show geographical disparities across Africa, with persistently low levels of attainment across the Sahel region, particularly in northern Nigeria, South Sudan and northern Kenya. In 2015, Ekiti state had the highest mean attainment in Nigeria among women of reproductive age, 11.3 years (95% uncertainty interval, 10.7–11.9) years, whereas many states in the northern region had averages below two years: Kebbi, 1.6 years (1.0–2.1); Yobe, 1.7 years (1.2–2.3); Sokoto, 1.5 years (1.0–2.1); and Zamfara, 1.6 years (1.1–2.2). For the same age range in Kenya, Nairobi province had the highest average attainment, 11.4 years (10.5–12.4), whereas the more rural North Eastern province had an average of 2.1 years (1.3–3.0). The lowest four regions across all of Africa had averages of less than 0.5 years, and all were rural regions in Chad: Daraba (0.5; 0.1–1.2), Kanem (0.4; 0.1–0.9), Barl El Gazal (0.4; 0.1–0.8) and Lac (0.4; 0.1–0.9). All outputs of these analyses at the national, first administrative subdivision (for example, state), second administrative

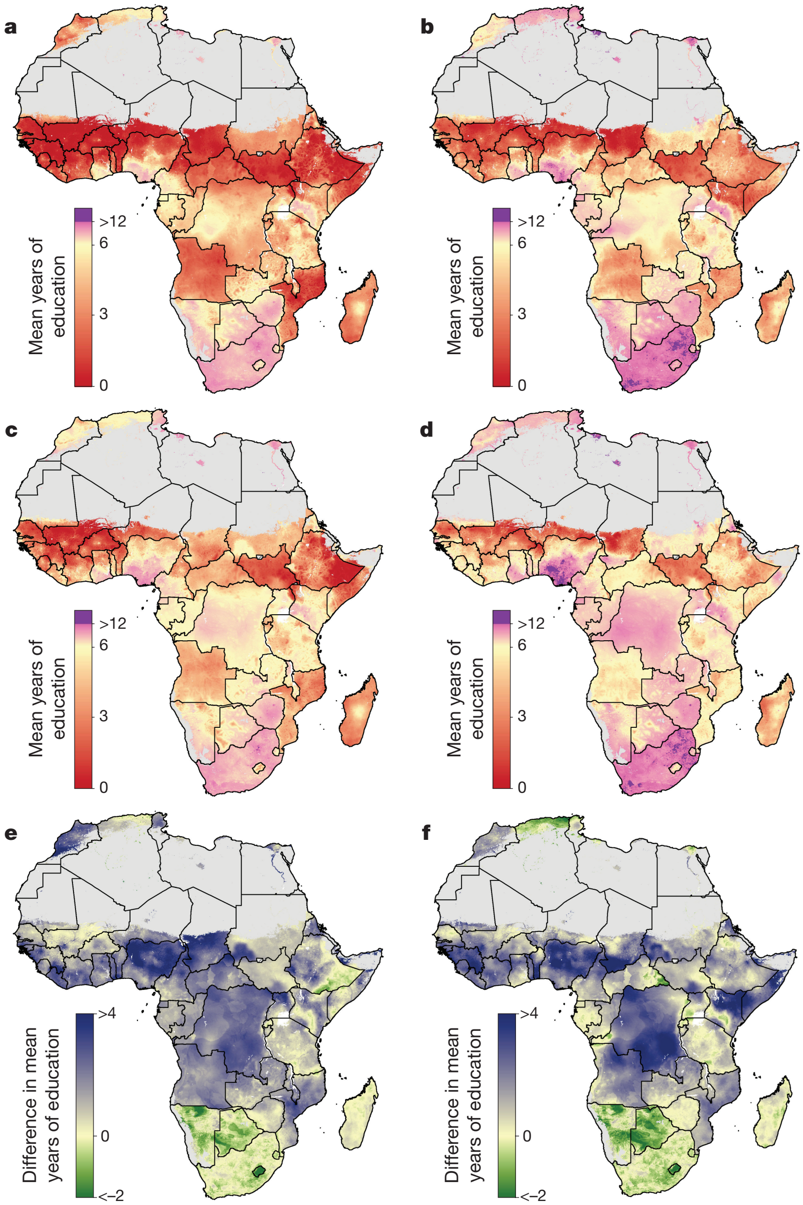


Figure 2 | Average educational attainment for and absolute difference between women and men aged 20–24 in 2000 and 2015. a–d, Average educational attainment for women (**a**, **b**) and men (**c**, **d**) aged 20–24 in 2000 (**a**, **c**) and 2015 (**b**, **d**). **e**, **f**, The absolute difference in average educational attainment between men and women aged 20–24 in 2000 (**e**) and 2015 (**f**). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[32,36–40].

subdivision (for example, district) and 5 × 5-km levels are publicly available from the Global Health Data Exchange (http://ghdx.healthdata.org/record/africa-educational-attainment-geospatial-estimates-2000-2015) and via bespoke data visualization tools (https://vizhub.healthdata.org/lbd/education).

Marked changes were observed over time when focusing on the 20–24 age range (Fig. 2a–d), with particular improvement observed in urban centres between 2000 and 2015 in Nigeria, Kenya, Ghana, Sudan and South Africa. Several populous urban states in Nigeria showed significant gains in average attainment for women since 2000, such as Abuja state, where attainment increased from 6.0 (4.7–7.2) to 9.7 years (9.0–10.5). In Ghana, the most highly educated urban regions in the southern part of the country demonstrated moderate increases in average attainment for women aged 20–24, such as Ashanti region, where attainment improved from 7.4 (6.9–7.9) to 9.9 years (9.5–10.4). Additionally, Ghana stands out in Western Africa for its improvements in more rural regions, for example, in the Northern region attainment improved from 1.8 (1.4–2.2) to 5.2 years (4.8–5.7) since 2000.

Implications for international goals

An explicit goal of SDG 4 is to eliminate sex-associated disparities across all levels of education by 2030[30]. We illustrate the gap in mean years of attainment between men and women for both age ranges (Figs 1e, f and 2e, f). Average attainment for men was significantly higher across the Sahel and Central Africa, particularly in the northern regions of Nigeria and Kenya that had very low levels of education in women of reproductive age (see Fig. 3). Here we use 'significantly' to refer to areas where 95% of the difference between Bayesian posterior predictive distributions was above zero (see Supplementary Information). These regions showed even stronger differences in the 20–24 age range, for which in some regions attainment in males was more than four years higher than in females (see Extended Data Fig. 1). Across states in 2015, we observed the largest difference in attainment by sex in the Kabia state of Chad, where men had achieved 5.8 more years (4.0–7.8) than women. In terms of statistical significance, 64 out of 77 states in Benin (representing 86% of the national population) had higher levels of attainment in males than females. The same was true for

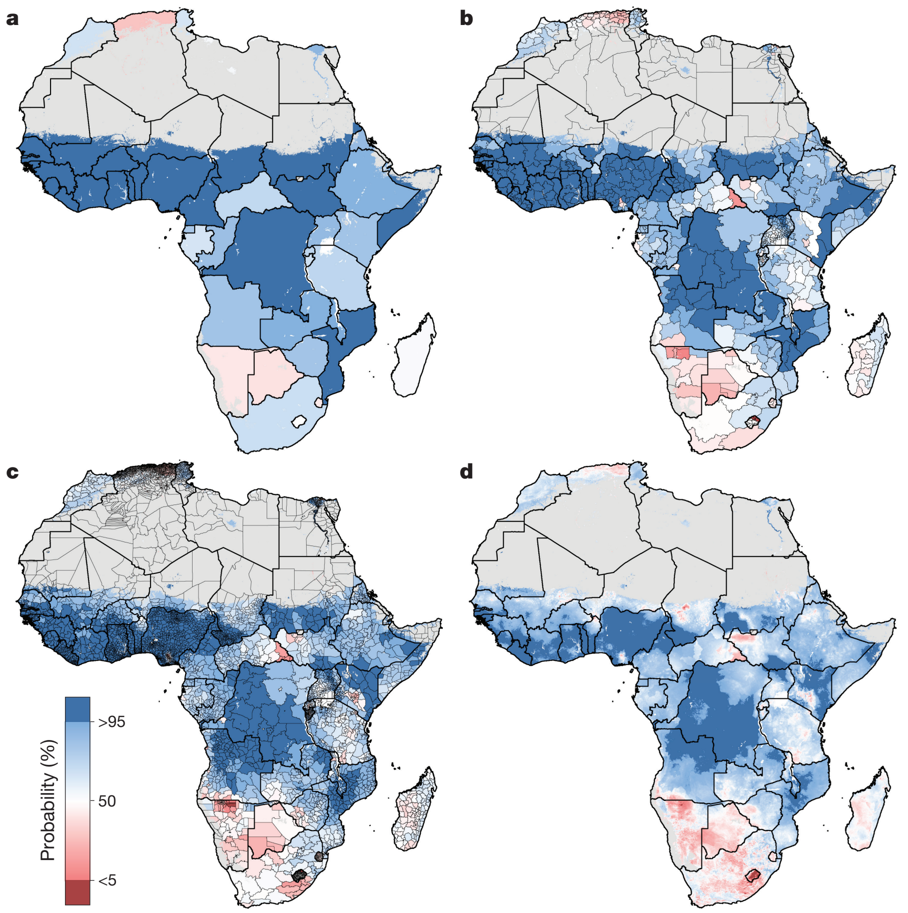


Figure 3 | Probability that male educational attainment is greater than female educational attainment for men and women aged 15–49 in 2015. **a**–**d**, Probabilities at the pixel level (**d**) were aggregated using 5 × 5-km resolution population data to the district level (**c**), province level (**b**) and national level (**a**). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[32,36–40].

all districts within Sierra Leone, Guinea, Guinea-Bissau and Togo. By contrast, average attainment trended towards higher levels for women across much of southern Africa in 2015; however, this difference was never significant. We observed no significant differences by sex for any district within South Africa, Botswana, Zimbabwe, Rwanda and others.

We further examined these trends in educational opportunity by applying a threshold for attainment. UNESCO defines basic education as completing the first nine years of formal schooling, including primary education (1–6 years of schooling) and lower secondary (7–9 years of schooling)[31]. The mean of 1,000 realizations of our full model is shown in Figs 1, 2. The Bayesian modelling framework that we used enables probabilistic inferences to be made about the likelihood that such targets have been met, on the basis of the confidence of the predictions (see Supplementary Information). In Figure 4, we illustrate the probability of average attainment being above six years in 2015 for women of reproductive age, or the equivalent of completing primary education (see Extended Data Fig. 2 for women aged 20–24). Despite SDG 4 not containing specific targets on years of attainment, this threshold was selected to highlight how substantial work remains in order to achieve even basic levels of education in many subnational regions within Africa.

We use high-resolution population data to aggregate these probabilities to different administrative levels for increased use in policy development and targeted intervention strategies, as well as to demonstrate the value of geospatial estimation for showing disparities within countries[32]. For instance, at the national level, the average woman of reproductive age in Nigeria has completed primary school in 2015. At lower geographical levels, however, these probabilities ranged from almost 0 to 100% of the population depending on the district or grid cells within the district (Fig. 3). Across Africa, many areas had averages that we could reasonably conclude were less than primary school completion (less than 5% probability of being greater than six years), but others were less certain. These regions may be less certain because our estimates were very close to six years, or because our estimates had wide uncertainty intervals (see Supplementary Information). Using the precision public health paradigm, these results have important implications for investment in education. Areas that were very unlikely (less than 5%) to be achieving primary school completion in 2015 should have investment aimed at improved access to basic education (examples of such measures are discussed in the Supplementary Discussion). Many areas with higher uncertainties probably not only have very low averages, but also require increased data collection efforts. This echoes the call in precision public health to invest in quality data at the local level to target interventions most equitably and efficiently[24].

Discussion, limitations and future work

This study represents a notable application of Bayesian geostatistical methods in a comprehensive, geolocated dataset to model educational attainment with refined spatial and temporal resolution. Our estimates show that although attainment has generally improved for women of reproductive age in Africa since 2000, these gains have now stagnated in many subnational regions. We also demonstrate that in 2015, gaps remain in attainment between the sexes in many areas across Africa; these gaps were relatively stable over time. These findings suggest that both men and women are experiencing progress in educational attainment, but the achievement of greater equity by sex remains out of reach for much of Africa.

Geographical inequality is only one form of inequality that can be used to investigate disparities below the national level. While our

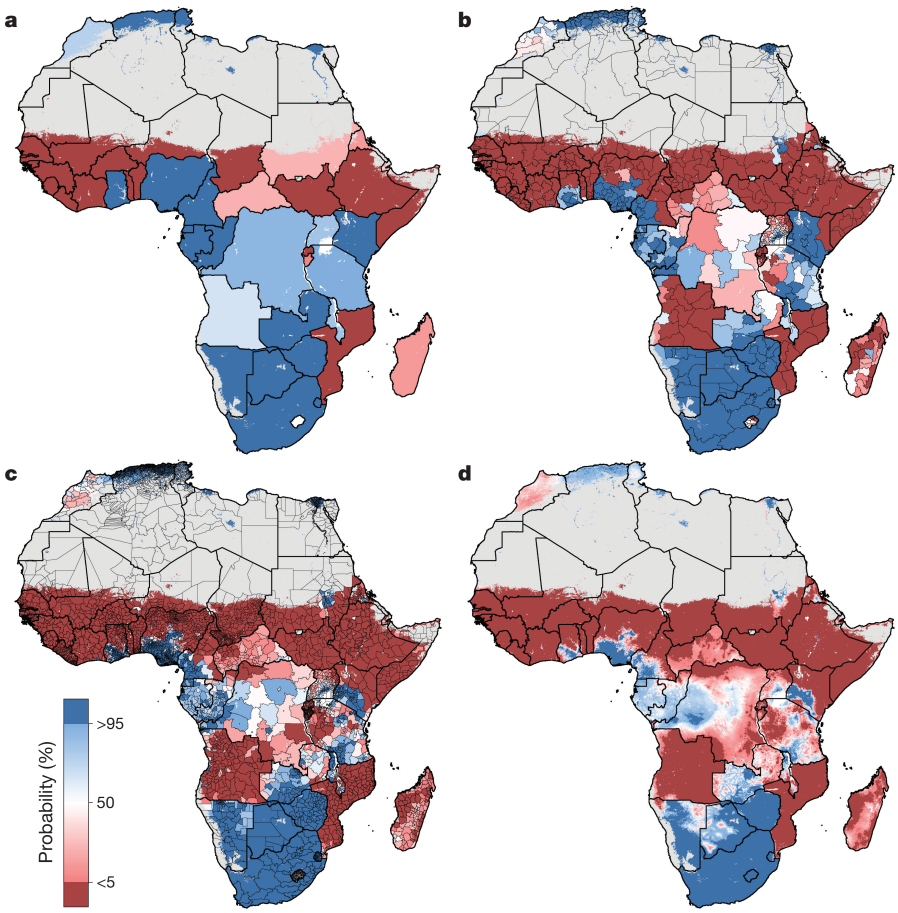


Figure 4 | Probability that average educational attainment is greater than six years in 2015 among women of reproductive age (15–49). **a**–**d**, Probabilities at the pixel level (**d**) were aggregated using 5 × 5-km resolution population data to the district level (**c**), province level (**b**) and national level (**a**). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[32,36–40].

framework allows us to explore geographical differences at a refined spatial level, there are many other dimensions that contribute to observed population inequities, such as social stratification by race, ethnicity or wealth (see Supplementary Discussion for limitations). Although further work is needed to explore additional forms of inequality, this predictive analysis has immediate relevance for policy development. First, our analysis maps a human capital indicator across Africa that is particularly relevant for the evolving global development agenda[33]. Second, and even more importantly, we are specifically considering educational attainment in women of reproductive age (and gender disparities in education) as a critical social determinant of maternal and thus child health[9–14].

Given the intersection between educational attainment for women of reproductive age and maternal and child health targets[34,35], these results have important implications for targeted investment to improve entrenched geographical and sex disparities. Communities with low education levels for women may be more likely to fail in public health interventions aimed at increasing prenatal care utilization, treatment adherence or contraception use[9,12,19,20]. Targeting precision health interventions without considering the landscape of human capital indexed by educational attainment poses sustainability risks, such as unrealistic assumptions about care-seeking behaviour and retention. In addition to the implications for health intervention, the global health agenda must also consider education and improved attainment as a goal itself in building sustainable, healthy populations.

Clearly the ultimate goal of SDG 4 extends beyond attainment to the quality of education. Nevertheless, as the global policy dialogue shifts to focusing on learning outcomes (see Supplementary Discussion), our results directly identify where gaps in basic education persist. These results can be used to improve accountability in need-based investment strategies from the national to local level. For communities which we have identified as having very low attainment, localized information can help to elucidate the drivers of low attendance and inform effective investment strategies.

Improving educational attainment among women of reproductive age has cross-cutting benefits for the SDG targets related to maternal and child health. This approach demonstrates the benefits of leveraging spatial information for modelling of human capital indicators in which data are correlated across space and time. This study emphasizes how documenting national-level trends in attainment masks pronounced variation across subnational areas. Despite progress, these findings suggest that large areas in sub-Saharan Africa still lag in meeting basic education targets, especially for women. In order to deliver on the promise of inclusive and equitable education for all[3], it is critical for investments in education to be informed by locally relevant information so that no community is left behind.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 August 2017; accepted 17 January 2018.

1. UNESCO. *Global Education Monitoring Report.* (UNESCO, 2016).
2. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development.* (United Nations, 2015).
3. United Nations. *Goal 4: Ensure Inclusive and Quality Education for All and promote Lifelong Learning.* (United Nations, 2016).
4. Klasen, S. *Does Gender Inequality Reduce Growth and Development: Evidence from Cross-Country Regressions.* Working Paper Series No. 7 (World Bank, 2000).

5. Klasen, S. Low schooling for girls, slower growth for all? Cross-country evidence on the effect of gender inequality in education on economic development. *World Bank Econ. Rev*. **16,** 345–373 (2002).
6. Klasen, S. & Lamanna, F. The impact of gender inequality in education and employment on economic growth: new evidence for a panel of countries. *Fem. Econ.* **15,** 91–132 (2009).
7. UNESCO. *Reducing Global Poverty through Universal Primary and Secondary Education*. Policy Paper 32 (UNESCO, 2017).
8. Abel, G. J., Barakat, B., Kc, S. & Lutz, W. Meeting the Sustainable Development Goals leads to lower world population growth. *Proc. Natl Acad. Sci. USA* **113,** 14294–14299 (2016).
9. Gakidou, E., Cowling, K., Lozano, R. & Murray, C. J. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *Lancet* **376,** 959–974 (2010).
10. Caldwell, J. C. How is greater maternal education translated into lower child mortality? *Health Transit. Rev*. **4,** 224–229 (1994).
11. Caldwell, J. C. Education as a factor in mortality decline: an examination of Nigerian data. *Popul. Stud. (NY)* **33,** 395 (1979).
12. Jejeebhoy, S. J. *Women's Education, Autonomy, and Reproductive Behaviour: Experience from Developing Countries* (Clarendon, 1995).
13. Desai, S. & Alva, S. Maternal education and child health: is there a strong causal relationship? *Demography* **35,** 71–81 (1998).
14. Basu, A. M. & Stephenson, R. Low levels of maternal education and the proximate determinants of childhood mortality: a little learning is not a dangerous thing. *Soc. Sci. Med.* **60,** 2011–2023 (2005).
15. Fuchs, R., Pamuk, E. & Lutz, W. Education or wealth: which matters more for reducing child mortality in developing countries? *Vienna Yearb. Popul. Res.* **8,** 175–199 (2010).
16. Boyle, M. H. *et al.* The influence of economic development level, household wealth and maternal education on child health in the developing world. *Soc. Sci. Med.* **63,** 2242–2254 (2006).
17. Pamuk, E. R., Fuchs, R. & Lutz, W. Comparing relative effects of education and economic resources on infant mortality in developing countries. *Popul. Dev. Rev*. **37,** 637–664 (2011).
18. Kravdal, Ø. Child mortality in India: the community-level effect of education. *Popul. Stud*. **58,** 177–192 (2004).
19. LeVine, R. A., LeVine, S., Schnell-Anzola, B., Rowe, M. L. & Dexter, E. *Literacy and Mothering: How Women's Schooling Changes the Lives of the World's Children* (Oxford Univ. Press, 2012).
20. Makate, M. & Makate, C. The causal effect of increased primary schooling on child mortality in Malawi: universal primary education as a natural experiment. *Soc. Sci. Med.* **168,** 72–83 (2016).
21. UNESCO. *Aid to Education Is Stagnating and Not Going to Countries Most in Need.* (UNESCO, 2017).
22. UNESCO. *Education 2030. World Educ. Forum 2015* (UNESCO, 2015).
23. UNESCO. *Is real progress being made in the equitable provision of education? #PISAresults.* http://www.iiep.unesco.org/en/real-progress-being-made-equitable-provision-education-pisaresults-3915 (IIEP UNESCO, 2017).
24. Dowell, S. F., Blazes, D. & Desmond-Hellmann, S. Four steps to precision public health. *Nature* **540,** 189–191 (2016).
25. Bosco, C. *et al.* Exploring the high-resolution mapping of gender-disaggregated development indicators. *J. R. Soc. Interface* **14,** 20160825 (2017).
26. Roberts, D. A. *et al.* Benchmarking health system performance across regions in Uganda: a systematic analysis of levels and trends in key maternal and child health interventions, 1990–2011. *BMC Med.* **13,** 285 (2015).
27. Gething, P. W. *et al.* Mapping *Plasmodium falciparum* mortality in Africa between 1990 and 2015. *N. Engl. J. Med.* **375,** 2435–2445 (2016).
28. Bhatt, S. *et al.* The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526,** 207–211 (2015).
29. Diggle, P. & Ribeiro, P. J. *Model-based Geostatistics* (Springer, 2007).
30. United Nations. *Sustainable Development Goals — United Nations* (United Nations, 2015).
31. UNESCO. *UNESCO Operational Definition Of Basic Education Thematic Framework* (UNESCO, 2007).
32. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4,** 170004 (2017).
33. The World Bank. *World Bank Group President Jim Yong Kim Speech at the 2017 Annual Meetings Plenary;* http://www.worldbank.org/en/news/speech/2017/10/13/wbg-president-jim-yong-kim-speech-2017-annual-meetings-plenary-session (2017).
34. Nilsson, M., Griggs, D. & Visbeck, M. Policy: map the interactions between Sustainable Development Goals. *Nature* **534,** 320–322 (2016).
35. Osgood-Zimmerman, A. *et al.* Mapping child growth failure in Africa between 2000 and 2015. *Nature* https://doi.org/10.1038/nature25760 (2018).
36. GeoNetwork. *Global Administrative Unit Layers (GAUL)*; http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691 (2015).
37. LP DAAC. Combined MODIS 5.1 dataset; available at: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12q1 (accessed 1 June 2017).
38. World Wildlife Fund. *Global Lakes and Wetlands Database Level 3* (2004); https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (accessed 1 June 2017).
39. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol. (Amst.)* **296,** 1–22 (2004).
40. WorldPop. WorldPop dataset; available at: http://www.worldpop.org.uk/data/get_data/ (accessed 7 July 2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grant OPP1132415 and OPP1093011 from the Bill & Melinda Gates Foundation, as well as grant OPP1070441 to E.G.

Author Contributions S.I.H., E.G. and N.G. conceived and planned the study. J.F. and J.L. obtained, extracted, processed and geopositioned educational attainment data. N.G., A.O.-Z. and R.B. wrote the computer code and designed and carried out the statistical analyses with input from R.C.R. N.G. and L.E. prepared tables and figures. L.E. and D.J.W. constructed covariate data layers. S.I.H., E.G., N.G., J.F., C.J.L.M., S.S.L., A.H.M., N.F., S.E.R., A.D., J.F.M., D.C.C., C.S., M.H.B. and A.J.L. provided intellectual inputs into aspects of this study. N.G., S.I.H. and E.G. wrote the first draft of the manuscript and all authors contributed to subsequent revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.I.H. (sihay@uw.edu) or E.G. (gakidou@uw.edu).

Reviewer Information *Nature* thanks E. Giorgi, W. Lutz and B. Reich for their contribution to the peer review of this work.

METHODS

Overview. Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). Using a Bayesian model-based geostatistical framework and synthesizing geolocated data from 173 household and census datasets, this analysis provides 5 × 5-km estimates of mean years of education for women of reproductive age (15–49), women aged 20–24, and equivalent male age-bins between 2000–2015 in Africa. This includes 48 countries in mainland Africa, as well as islands for which we had survey data, including Madagascar, Comoros, and São Tomé and Príncipe. We did not estimate for Mauritius, Seychelles or Cape Verde, as no available survey data could be sourced. Analytical steps are described below and additional detail can be found in the Supplementary Information.

Data. We compiled a database of 173 survey and census datasets in Africa that contained geocoding of subnational administrative boundaries or precise coordinates for sampled clusters. These included datasets from the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS) and Integrated Public Use Microdata Series (IPUMS)[41–43] (see Supplementary Table 2). We extracted demographic, education and sample design variables. The coding of educational attainment varies across survey families. In many surveys, respondents can indicate their level of attainment on a continuous year scale. In others, respondents may only have several aggregate categories such as 'Secondary completion', 'Primary completion', or 'less than primary'. When all that is known is that an individual completed a particular level of education, but it is not known if they continued onto the next level, a theoretical level of completion must be assigned to the individual in order to estimate summary statistics for the population such as mean years of educational attainment. For example, if the option 'Primary completion' (6 years) is followed by 'Secondary completion' (12 years), it can be assumed that an individual who only selects the former has attained between 6 and 12 years of education. In previous literature examining trends in mean years of education, the assumption is made that all of these individuals have 6 years, or sometimes the midpoint of the feasible range (9)[44,45]. Trends in the single-year data demonstrate that this assumption introduces compositional bias in the estimation of attainment trends over time and space, as differences in true drop-out patterns or binning schema could lead to biased mean estimates.

For this analysis, we used a recently developed method that selects a training subset of similar surveys across time and space to estimate the true single-year distribution of binned datasets (J.F., N.G. & E.G., manuscript in preparation). This algorithmic approach markedly reduces bias in summary statistics estimated from datasets with binned coding schemes. The years in all coding schemes were mapped to the country- and year-specific references in the UNESCO International Standard Classification of Education (ISCED) for comparability[46]. We used a top coding of 18 years on all data; this is a common threshold in many surveys that have a cap and it is reasonable to assume that the importance of education for health outcomes (and other related SDGs) greatly decreases after what is the equivalent of 2 to 3 years of graduate education in most systems.

Data were aggregated to mean years for women of reproductive age (15–49) to measure progress towards the SDG 4 target[2]. A subset of the data for a smaller age range of women aged 20–24 was also examined to track temporal shifts as well as the effects of large educational initiatives in Africa since 2000. Equivalent age-bins were aggregated for males in order to examine differences in mean years of attainment by sex. Where precise coordinates were available, data were aggregated to mean years at a specific latitude and longitude assuming a simple random sample, as the cluster is the primary sampling unit for the stratified design of all DHS and MICS surveys. Where only geography information was available at the level of administrative units, data were aggregated according to their sample design. For aggregation to administrative units for which the survey was not sampled to be representative, design effects were re-estimated using a package for analysing complex survey data in R[47].

Spatial covariates. In order to leverage strength from locations with observations to the entire spatiotemporal domain, we compiled several 5 × 5-km raster layers of possible socio-economic and environmental correlates of education in Africa (see Supplementary Table 3 and Supplementary Fig. 5). Acquisition of temporally dynamic datasets, where possible, was prioritized in order to best match our observations and thus predict the changing dynamics of educational attainment. Of the 29 covariates included, 23 were temporally dynamic. The remaining six covariate layers were temporally static, and were applied uniformly across all modelling years. More information, including plots of all covariates, can be found in the Supplementary Information.

Our primary goal is to provide educational attainment predictions across the African continent at a high resolution and we have used methods to provide the best out-of-sample predictive performance at the expense of inferential understanding. In order to select covariates and capture possible nonlinear effects and complex interactions between them, an ensemble covariate modelling method was implemented[48]. For each region three sub-models were fit to our dataset using all of our covariate data as explanatory predictors: generalized additive models, boosted regression trees and lasso regression. Each sub-model was fit using fivefold cross-validation to avoid overfitting and the out-of-sample predictions from across the five holdouts are compiled into a single comprehensive set of predictions from that model. Additionally, the same sub-models were also run using 100% of the data and a full set of in-sample predictions were created. The five sets of out-of-sample sub-model predictions were fed into the full geostatistical model as the explanatory covariates when performing the model fit. The in-sample predictions from the sub-models were used as the covariates when generating predictions using the fitted full geostatistical model. This methodology maximizes out-of-sample predictive performance at the expense of no longer being able to provide statistical inferences on causality. A recent study has shown that this ensemble approach can improve predictive validity by up to 25% over an individual model[48]. More details on this approach can be found in the Supplementary Information.

Analysis. *Geostatistical model.* Gaussian data are modelled within a Bayesian hierarchical modelling framework using a spatially and temporally explicit hierarchical generalized linear regression model to fit mean years of education attainment in five regions in Africa as defined in the Global Burden of Diseases, Injuries, and Risk Factors (GBD) study[49] ('Northern', 'Western', 'Southern', 'Central' and 'Eastern'; see Extended Data Fig. 3). GBD study design sought to create regions on the basis of two primary criteria: epidemiological homogeneity and geographical contiguity[49]. For each GBD region, we approximated the posterior distribution of our Bayesian model:

$$\mathrm{edu}_i|\mu_i, \tau, s_i \sim \mathrm{Gaussian}(\mu_i, \tau, s_i)$$

$$f_{\mathrm{edu}_i|\mu_i,\,\tau,\,s_i}(\mathrm{edu}_i) = \frac{\sqrt{\tau s_i}}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\,\tau s_i(\mathrm{edu}_i - \mu_i)^2\right)$$

$$\mu_i = \beta_0 + \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_{\mathrm{GP}i} + \epsilon_i$$

$$\epsilon_i \sim N\left(0, \sigma^2_{\mathrm{nug}}\right)$$

$$\boldsymbol{\epsilon}_{\mathrm{GP}}|\Sigma_{\mathrm{space}}, \Sigma_{\mathrm{time}} \sim \mathrm{GP}(0, \Sigma_{\mathrm{space}} \otimes \Sigma_{\mathrm{time}})$$

$$\Sigma_{\mathrm{space}} = \frac{2^{1-\nu}}{\tau \times \Gamma(\nu)} \times (\kappa\boldsymbol{D})^{\nu} \times \mathrm{K}_{\nu}(\kappa\boldsymbol{D})$$

$$\Sigma_{\mathrm{time}_{j,k}} = \rho^{|t_k - t_j|}$$

We model the mean years of attainment at cluster *i* as Gaussian data given precision τ and a fixed scaling parameter s_i. We use the sample size in each cluster as our scaling parameter. We have suppressed the notation, but the means (edu_i), scaling parameters (s_i), predictions from the three submodels ($\boldsymbol{X}_i$), and residual terms (ϵ_*) are all indexed at a space–time coordinate. The means (edu_i) represent an individual's expected educational attainment given that they live at that particular location. Mean attainment was modelled as a linear combination of the three sub-models (GAM, BRT and lasso), $\boldsymbol{X}_i$, a correlated spatiotemporal error term, $\epsilon_{\mathrm{GP}i}$, and an independent nugget effect, ϵ_i. Coefficients, $\boldsymbol{\beta}$, on the sub-models represent their respective predictive weighting on the mean, while the joint error term, $\boldsymbol{\epsilon}_{\mathrm{GP}}$, accounts for residual spatiotemporal autocorrelation between individual data points that remains after accounting for the predictive effect of the sub-model covariates and the nugget, ϵ_i, is an independent error term. The residuals, $\boldsymbol{\epsilon}_{\mathrm{GP}}$, are modelled as three-dimensional Gaussian processes in space–time centred at zero and with a covariance matrix constructed from a Kroenecker product of spatial and temporal covariance kernels. The spatial covariance, Σ_{space}, is modelled using an isotropic and stationary Matérn function[50], and temporal covariance, Σ_{time}, as an annual autoregressive (AR1) function over the 16 years represented in the model. This approach leveraged the data's residual correlation structure to more accurately predict attainment estimates for locations with no data, while also propagating the dependence in the data through to uncertainty estimates[51]. The posterior distributions were fit using computationally efficient and accurate approximations in R INLA (integrated nested Laplace approximation) with the stochastic partial differential equations approximation to the Gaussian process residuals[52]. Pixel-level uncertainty intervals were generated from 1,000 draws (that is, statistically plausible candidate maps)[53] created from the posterior-estimated distributions of modelled parameters.

To transform pixel level estimates into a range of information useful to a wide constituency of potential users, these estimates were aggregated from the 1,000 candidate maps up to district, provincial and national levels using

5 × 5-km population data[32]. This aggregation also enabled the calibration of estimates to national GBD estimates for 2000, 2005, 2010 and 2015. This was achieved by calculating the ratio of the posterior mean national-level estimate from each candidate map draw in the analysis to the posterior mean national estimates from GBD, and then multiplying each cell in the posterior sample by this ratio. This method also enabled the incorporation of the calibration into the pixel level uncertainties and thus to the uncertainties at the different levels of aggregation. The median raking factors for women aged 15–49, men aged 15–49, women aged 20–24 and men aged 20–24 were 0.926 (interquartile range (IQR): 0.794–1.084), 0.895 (IQR: 0.761–1.012), 1.036 (IQR: 0.798–1.031) and 1.053 (IQR: 0.861–1.233), respectively, indicating close agreement with GBD estimates. Scatter plots comparing national level estimates from this analysis with GBD estimates can be found in Supplementary Figs 24–27.

Although the model can predict at all locations covered by available raster covariates, all final model outputs for which land cover was classified as 'barren or sparsely vegetated' were masked, on the basis of the most recently available MODIS satellite data (2013), as well as areas where the total population density was less than ten individuals per 1 × 1-km pixel in 2015. This step has led to improved understanding when communicating with data specialists and policy makers.

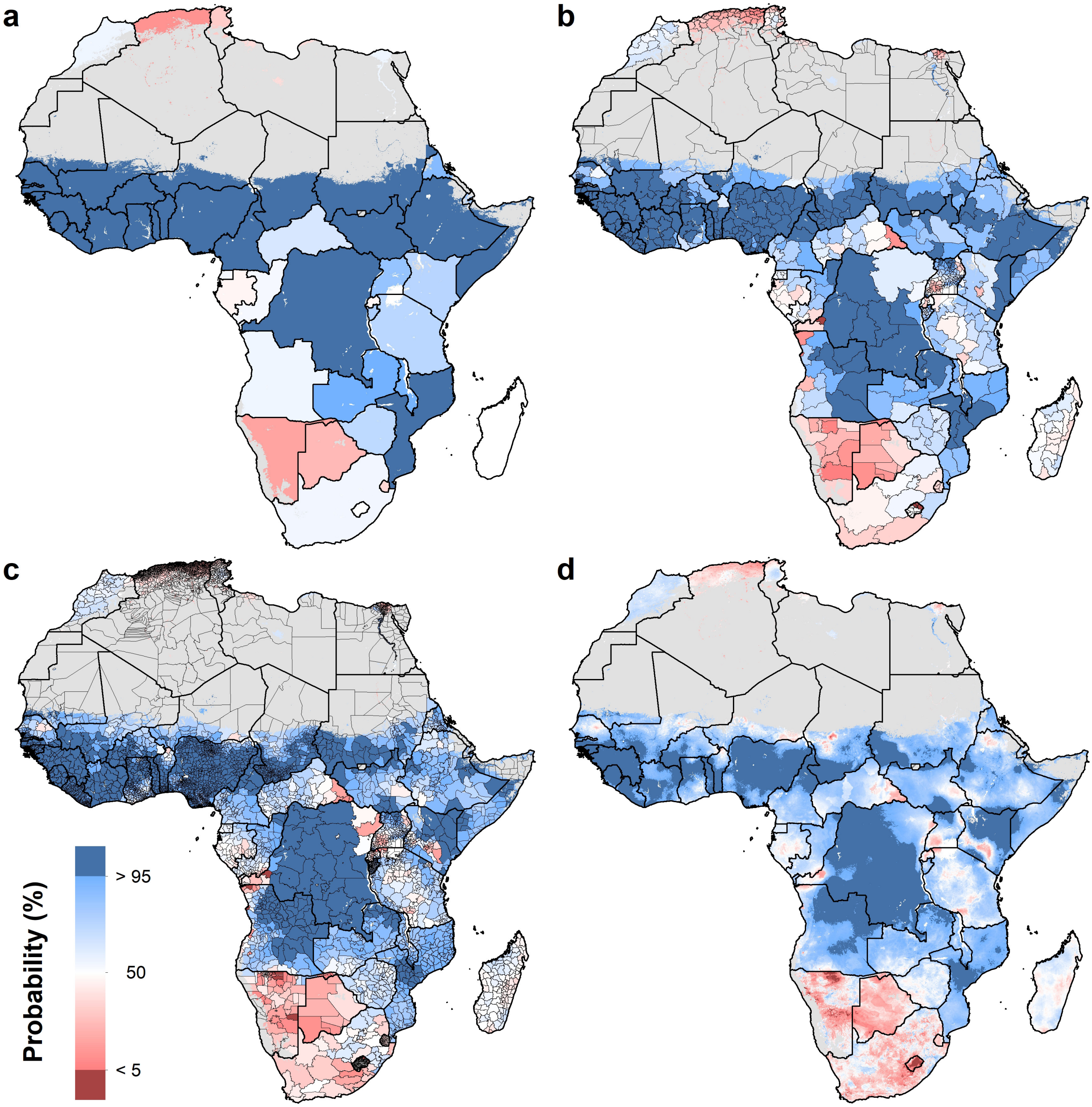
Model validation. Models were validated using spatially stratified fivefold out-of-sample cross-validation. In order to offer a more stringent analysis by respecting some of the spatial correlations in the data, holdout sets were created by combining sets of spatially contiguous data. Validation was performed by calculating bias (mean error), total variance (root-mean-square error) and 95% data coverage within prediction intervals, and correlation between observed data and predictions. All validation metrics were calculated on the out-of-sample predictions from the fivefold cross-validation. Where possible, estimates from these models were compared against other existing estimates. Furthermore, measurements of spatial and temporal autocorrelation pre- and post-modelling were examined to verify correct recognition, fitting and accounting for the complex spatiotemporal correlation structure of the data. All validation procedures and corresponding results are provided in the Supplementary Information.

Code availability. All code used for these analyses is available online at http://ghdx.healthdata.org/record/africa-educational-attainment-geospatial-estimates-2000-2015.

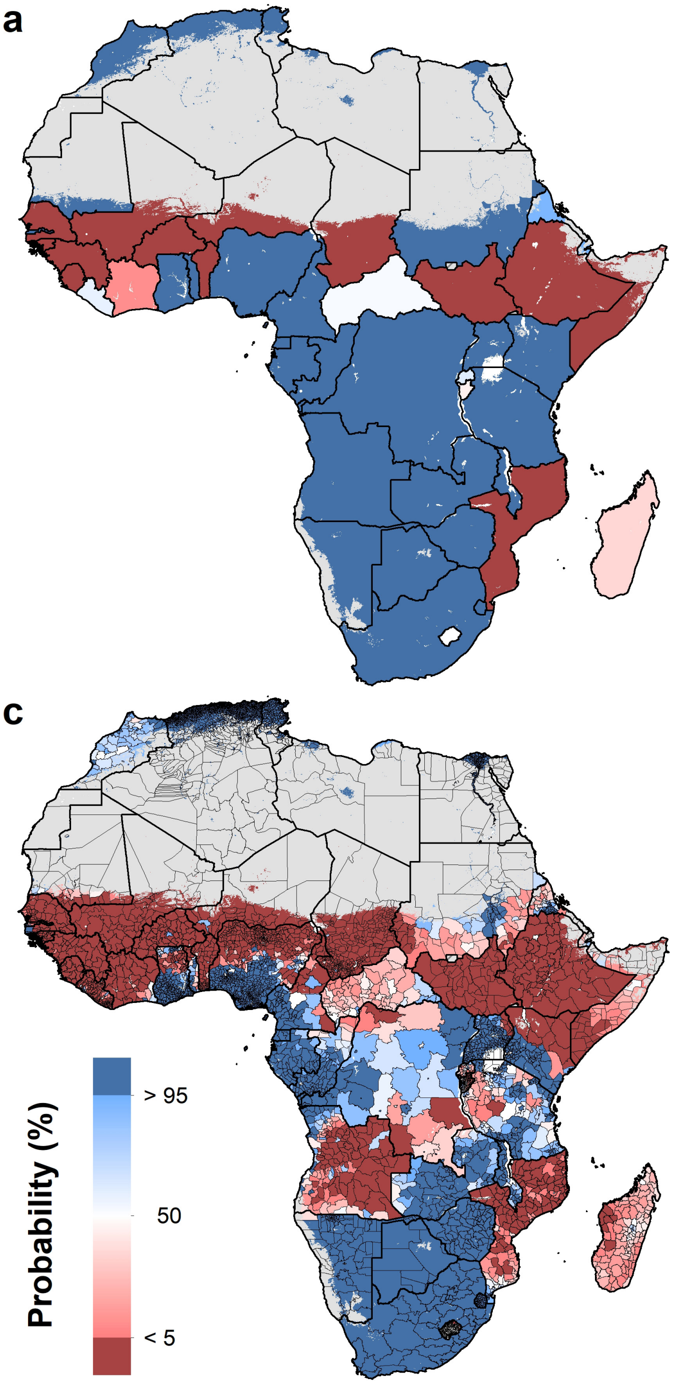
Data availability. The findings of this study are supported by data that are available from public online repositories, data that are publicly available upon request of the data provider and data that are not publicly available because of restrictions by the data provider, which were used under license for the current study, but may be available from the authors upon reasonable request and permission of the data provider. A detailed table of data sources and availability can be found in Supplementary Table 2.

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects[36]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota[37]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel[38,39]. Populations were retrieved from WorldPop[32,40].

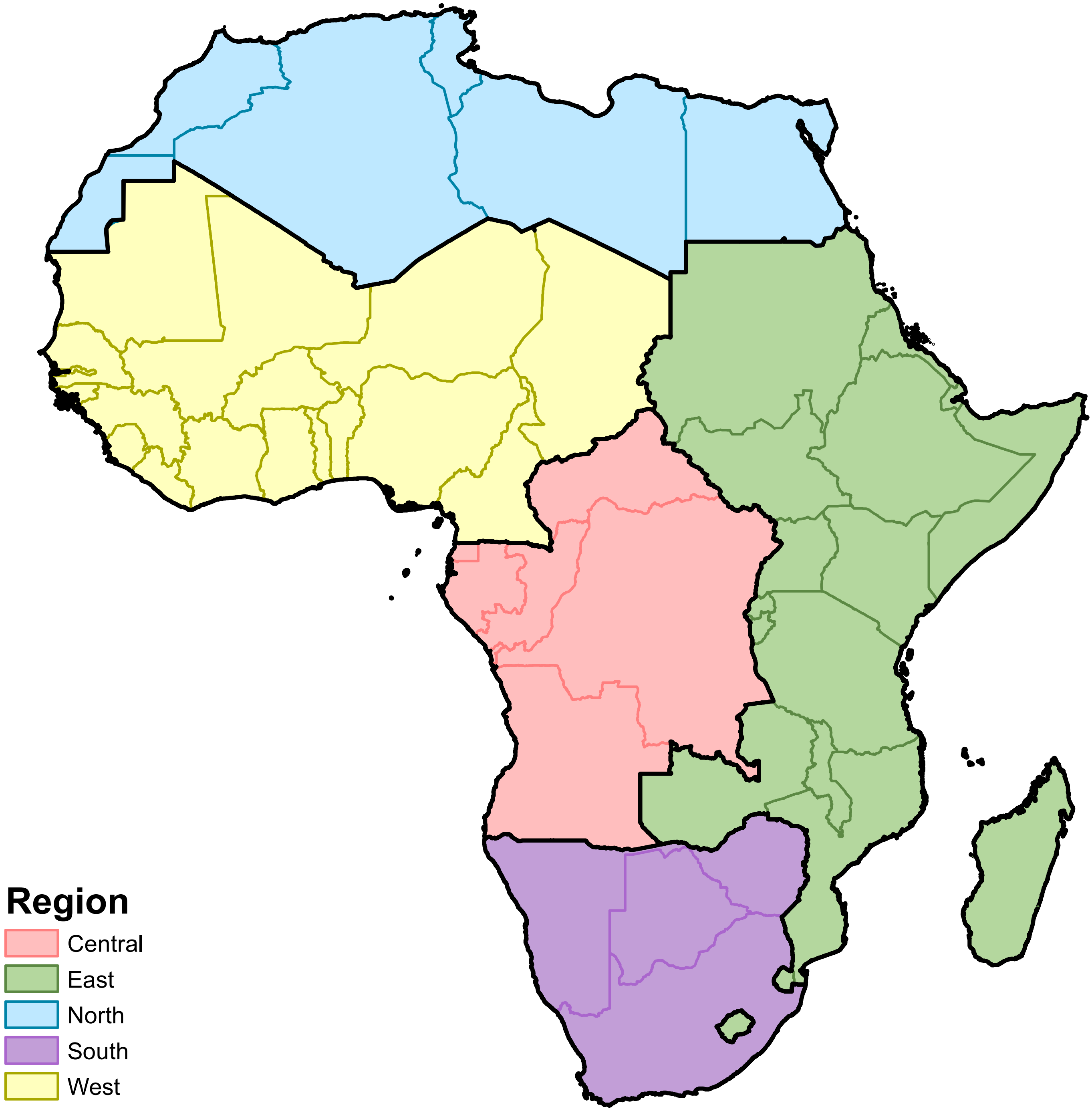
41. ICF. The DHS Program, Data. http://dhsprogram.com/data/ (1998).
42. UNICEF. *Multiple Indicator Cluster Survey (MICS).* https://www.unicef.org/statistics/index_24302.html (UNICEF, 2010).
43. Minnesota Population Center. *IPUMS International*; https://international.ipums.org/international (2015).
44. Barro, R. J. & Lee, J.-W. *International Comparisons of Educational Attainment.* NBER Working Paper No. 4349 (NBER, 1993).
45. Barro, R. J. & Lee, J.-W. *A New Data Set of Educational Attainment in the World, 1950–2010.* NBER Working Paper No. 15902 (NBER, 2010).
46. UNESCO. ISCED Mappings; http://uis.unesco.org/en/isced-mappings (UNESCO, 2016).
47. Lumley. T. *Complex Surveys: A Guide to Analysis Using R* (Wiley-Blackwell, 2014).
48. Bhatt, S. *et al.* Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalisation. *J. R. Soc. Interface* **14,** 20170520 (2016).
49. Murray, C. J. *et al.* GBD 2010: design, definitions, and metrics. *Lancet* **380,** 2063–2066 (2012).
50. Stein, M. L. *Interpolation of Spatial Data* (Springer, 1999).
51. Waller, L. & Carlin, B. in *Handbook of Spatial Statistics* (eds Gelfand, A. *et al.*) 217–243 (CRC, 2010).
52. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71,** 319–392 (2009).
53. Patil, A. P., Gething, P. W., Piel, F. B. & Hay, S. I. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol.* **27,** 246–253 (2011).
54. GBD 2015 DALYs and HALE Collaborators. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388,** 1603–1658 (2016).



Extended Data Figure 1 | Probability that educational attainment in men is greater than attainment in women for men and women aged 20–24 in 2015. a–d, Probabilities at the pixel level (**d**) were aggregated using 5 × 5-km resolution population data to the district level (**c**), province level (**b**) and national level (A). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as ‘barren or sparsely vegetated’ are coloured in grey[32,36–40].



Extended Data Figure 2 | Probability that average educational attainment is greater than six years in 2015 among women aged 20–24. Probabilities at the pixel level (**d**) were aggregated using 5 × 5-km resolution population data to the district level (**c**), province level (**b**) and national level (A). Maps reflect administrative boundaries, land cover, lakes and population; pixels with fewer than ten people per 1 × 1 km and classified as 'barren or sparsely vegetated' are coloured in grey[32,36–40].



Extended Data Figure 3 | Map of modelling regions. Modelling regions were defined as the five GBD regions of Central (central sub-Saharan Africa), East (eastern sub-Saharan Africa), North (North Africa and the Middle East), South (Southern sub-Saharan Africa) and West Africa (Western sub-Saharan Africa)[54]. As this study was limited to mainland Africa and African island nations, select countries were excluded from the North Africa and Middle East region (Afghanistan, Bahrain, Iran, Iraq, Jordan, Kuwait, Lebanon, Oman, Palestinian territories, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates and Yemen). Western Sahara was included as part of the North region. Several countries were moved to East (Lesotho and Swaziland from South, Sudan from North) to make high-income status more similar in the North and South regions.

doi:10.1038/nature25741

Population snapshots predict early haematopoietic and erythroid hierarchies

Betsabeh Khoramian Tusi[1]*, Samuel L. Wolock[2]*, Caleb Weinreb[2]*, Yung Hwang[1], Daniel Hidalgo[1], Rapolas Zilionis[2], Ari Waisman[3], Jun R. Huh[4], Allon M. Klein[2] & Merav Socolovsky[1,5]

The formation of red blood cells begins with the differentiation of multipotent haematopoietic progenitors. Reconstructing the steps of this differentiation represents a general challenge in stem-cell biology. Here we used single-cell transcriptomics, fate assays and a theory that allows the prediction of cell fates from population snapshots to demonstrate that mouse haematopoietic progenitors differentiate through a continuous, hierarchical structure into seven blood lineages. We uncovered coupling between the erythroid and the basophil or mast cell fates, a global haematopoietic response to erythroid stress and novel growth factor receptors that regulate erythropoiesis. We defined a flow cytometry sorting strategy to purify early stages of erythroid differentiation, completely isolating classically defined burst-forming and colony-forming progenitors. We also found that the cell cycle is progressively remodelled during erythroid development and during a sharp transcriptional switch that ends the colony-forming progenitor stage and activates terminal differentiation. Our work showcases the utility of linking transcriptomic data to predictive fate models, and provides insights into lineage development *in vivo*.

The abundance of erythroid progenitors in haematopoietic tissue provides a unique opportunity to dissect how multipotent progenitors (MPPs) differentiate into a single lineage *in situ*, a process of fundamental biological interest and of clinical relevance. Erythropoiesis has two principal phases: erythroid terminal differentiation (ETD), in which GATA1-driven transcription remodels erythroid precursors into red blood cells through several well-described stages[1–3]; and an earlier, less-well-delineated phase of early erythropoiesis, in which haematopoietic stem cells (HSCs) differentiate through poorly defined intermediates into erythroid progenitors. Erythroid progenitors have been identified by their colony-forming potential in semi-solid medium, as either burst-forming (BFU-e) or colony-forming (CFU-e) progenitors[4,5]. A direct, complete and high-purity isolation of adult mouse BFU-e and CFU-e from haematopoietic tissue has not, to our knowledge, been attained[6–9]. More broadly, there are no known strategies that systematically identify the entire cellular and molecular trajectory of the early erythroid lineage as it first arises from the HSC compartment and progresses to the point at which ETD is activated.

Probing the earliest stages of erythropoiesis requires exploring how MPPs diversify into progenitors for each of the haematopoietic lineages. Single-cell approaches have recently challenged established models of haematopoiesis, showing that progenitor populations that were thought to be similar in their developmental stage and fate potentials are in fact highly heterogeneous in both respects[10–17]. Alternative models replaced the classic haematopoietic tree[18–20] with a 'flatter' hierarchy, in which unilineage progenitors derive directly from a heterogeneous set of lineage-biased multipotent progenitors. These new models are highly dependent on the tools used in the analysis of high-dimensional cell transcriptional states, which are currently undergoing intense innovation[21–26]. Descriptions of the haematopoietic hierarchy structure have so far relied on clustering[12], which may fail to capture continuums; diffusion maps[26,27], which are powerful for branching models but provide less detail of highly complex processes; and the ordering of progenitors on the basis of their similarity to differentiated cell types[16], which may overlook progenitors that do not resemble mature cells. Fluorescence-activated cell sorting (FACS) may also introduce bias into reconstructed developmental trajectories, through overly restrictive gates[15] and the loss of sensitive cells[28]. It is still not clear how to reconcile cell fate assays with the cell-state maps proposed from the results of single-cell profiling, and this remains a general challenge in stem-cell biology.

Here we investigated the derivation and subsequent development of erythroid progenitors by undertaking single-cell RNA sequencing (scRNA-seq) of a broad set of haematopoietic progenitors, using the inDrops platform[29]. We developed an analytical tool, population balance analysis (PBA), which can be used to predict cell fate probabilities from static snapshots of single-cell transcriptomes through dynamic inference. PBA allowed us to define a FACS strategy to isolate cells in progressive stages of early erythropoiesis. Using single-cell fate assays, we then confirmed a number of detailed predictions regarding the early haematopoietic hierarchy and erythroid developmental progression. The insights obtained into early erythropoietic fate control may be applicable to other differentiation models, and include novel erythropoietic regulators with potential therapeutic relevance.

scRNA-seq of Kit$^+$ progenitors

We performed scRNA-seq on Kit$^+$ haematopoietic progenitor cells (HPCs) isolated using magnetic beads from the bone marrow of adult mice (Fig. 1a). Kit is expressed on all haematopoietic stem and early progenitor cells[30,31], enabling an inclusive approach that preserves the relative abundance of progenitor cell states. After filtering, we carried forward 4,763 HPC transcriptomes for analysis (see 'Data availability' in Methods for interactive tools).

We visualized the scRNA-seq data using the SPRING algorithm[32] (Fig. 1b), which generates a graph of cells (graph nodes) connected to their nearest neighbours in gene expression space and projected into two dimensions using a force-directed graph layout. This visualization

[1]Department of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [2]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [3]Institute for Molecular Medicine, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany. [4]Division of Immunology, Department of Microbiology and Immunobiology and Evergrande Center for Immunological Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts, USA. [5]Department of Pediatrics, Hematology/Oncology Division, University of Massachusetts Medical School, Worcester, Massachusetts, USA.
*These authors contributed equally to this work.

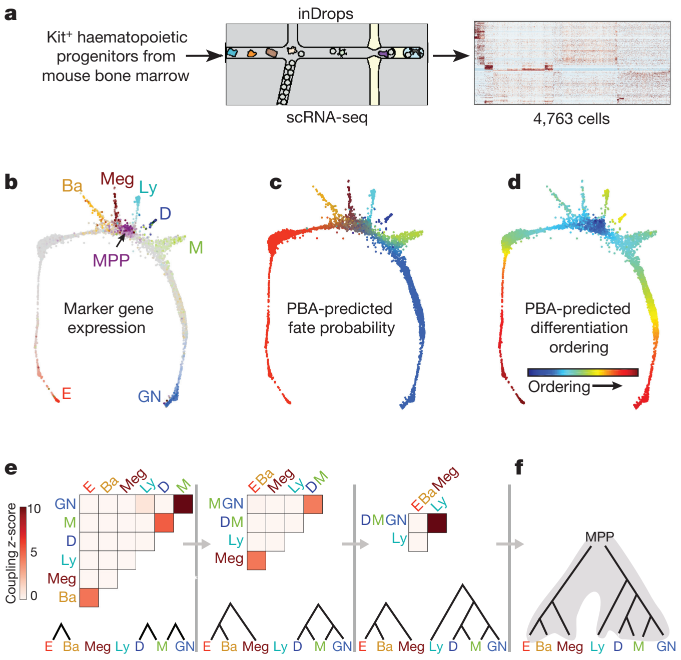


Figure 1 | The early haematopoietic hierarchy predicted by scRNA-seq. **a**, Schematic for scRNA-seq of Kit$^+$ mouse bone marrow. **b**, SPRING plot of single-cell transcriptomes. Each point is one cell. Colours indicate lineage-specific gene expression. Ba, basophilic or mast cell; D, dendritic; E, erythroid; GN, granulocytic neutrophil; Ly, lymphocytic; M, monocytic; Meg, megakaryocytic; MPP, multipotential progenitors. **c**, **d**, Parameterization of the cell-state graph using PBA, encoding the graph position of each cell by a set of predicted fate probabilities (**c**) (colours as in **b**), and pseudo-temporal ordering with MPPs at the origin, terminating with the most mature cells observed for each lineage (**d**). **e**, **f**, A cell-state hierarchy encodes the cell-graph topology. **e**, Lineage-biased states were identified by comparing the fraction of cells with PBA-predicted bilineage coupling with expected values from fate randomization. **f**, Iteratively joining fates on the basis of pairwise coupling revealed the cell-state hierarchy.

suggested that HPCs occupy a continuum of transcriptional states, rather than discrete metastable states, a result that contrasts with single-cell data from mature blood lineages and is supported by formal tests of graph interconnectivity (Extended Data Fig. 1a). When SPRING plots were coloured on the basis of the expression of lineage-specific markers (Fig. 1b, Supplementary Table 1), the cells were found to organize around an undifferentiated core, from which seven distinct branches emerge, corresponding to progenitors of the basophil or mast cell (Ba/Mast), granulocytic neutrophil, monocytic, dendritic, lymphoid, megakaryocytic and erythroid lineages. Although this structure depends critically on collecting cells using a broad selection marker, SPRING visualization of data from previous studies[12,15] revealed the same lineage relationships (Extended Data Fig. 1b).

PBA of the HPC continuum

To understand the differentiation trajectories that cells might follow, we developed the PBA[33] approach for studying single-cell continua (Extended Data Fig. 1c, d). PBA maps each cell to a low-dimensional space that encodes the cell-graph topology in the form of predicted cell fate probabilities. The derivation and limitations of PBA have been detailed elsewhere[33]. The core of PBA can be understood as the reconstruction of the memoryless stochastic dynamics of cells, which, through ongoing cell turnover, explains the observed steady-state distribution of cell states. PBA approximates the dynamics of cells as following the gradient of a potential landscape, which itself can be inferred through an asymptotic relationship between diffusion–drift processes and the spectral properties of the SPRING graph[33].

The probabilities obtained by PBA represent formal biophysical predictions for the fate of cells under simplified assumptions, but they can also be treated heuristically as encoding graph distances. Applied to our data, PBA defines seven putative commitment probabilities for each haematopoietic progenitor (Fig. 1c, Extended Data Fig. 1e), as well as the distance from the undifferentiated CD34$^{\mathrm{high}}$Sca1$^{\mathrm{high}}$ MPPs (Fig. 1d).

The transcriptional continuum of HPCs is hierarchical

We used PBA-predicted commitment probabilities to compute a coupling score that reflects whether any two fate potentials occur concurrently in single progenitors at higher rates than would be expected by chance (Fig. 1e). A transcriptional-state hierarchy was formalized by identifying correlated pairs of terminal fates and joining them iteratively until a multipotent state was reached (Fig. 1e, f). The resulting topology firmly supports the hierarchical view of haematopoiesis, with MPPs diverging into progenitors with either correlated erythroid, Ba/Mast and megakaryocytic fates, or with correlated lymphoid and myeloid fates (Fig. 1f). However, the transcriptional-state hierarchy emerges from correlations on a continuum, rather than from discrete populations. Additionally, it predicts two refinements over current models. First, the erythroid fate is correlated with the Ba/Mast fates. Second, among myeloid progenitors we identified dendritic–monocyte and granulocytic–monocyte coupling, but no dendritic–granulocytic coupling, suggesting that monocyte differentiation may occur through two distinct trajectories, a prediction that was very recently independently confirmed[34]. The PBA-formalized HPC hierarchy also allowed us to identify genes for which expression closely correlates with each cell fate choice (Extended Data Fig. 2, Supplementary Table 2).

Isolation of putative erythroid progenitors

To test PBA predictions (Fig. 1e–f), we developed a FACS strategy that isolates haematopoietic subpopulations defined by scRNA-seq. Guided by the single-cell expression patterns, we combined Kit expression with CD55, a marker of megakaryocytic and erythroid bias[10]; Kit$^+$CD55$^+$ cells were divided into subpopulations (P1–P5), using CD49f (encoded by *Itga6*) and megakaryocytic and erythroid markers[6,9,10] (Fig. 2a). Using reverse transcription with quantitative PCR (qRT–PCR) (Extended Data Fig. 3a), and scRNA-seq (11,241 cells post-filter) (Fig. 2b, Extended Data Fig. 3b–e), we mapped cells from each of the sorted subpopulations back to regions of the SPRING graph. We found that P1 and P2 represent high-purity subpopulations on the putative erythroid branch, with P1 predicted to be committed, and P2 mostly committed, to the erythroid fate (Fig. 1c). P3 and P4 are enriched for the Ba/Mast and megakaryocytic branches, respectively; P3 bifurcates into separate basophil and mast cell branches. P5 contains erythroid-biased oligopotent and multipotent cells. Myeloid, lymphoid and some MPP states are within the CD55$^-$ region of the plot.

Functional identification of correlated cell fates

We next examined the differentiation potential of the sorted P1–P5 populations, and by extension, the predicted fate probabilities for their corresponding transcriptional states. Colony-forming assays showed that P1 and P2 contain all of the unipotential erythroid progenitors and no other progenitors (Fig. 2c–e). P1 colonies were small and unifocal, maturing on day 3 or later (CFU-e) (Fig. 2c), whereas P2 colonies were largely multifocal, maturing on day 4 or later (BFU-e) (Fig. 2d). Thus, P1 is closer to erythroid maturation than P2, consistent with PBA predictions (Figs 1c, 2b). Furthermore, the transcriptional state of progenitors as defined by the SPRING map determines their ability to form either multifocal or unifocal colonies. Consistent with the SPRING plot, the less-differentiated P5 population gave rise to mixed myeloid colonies, and P4 was enriched for megakaryocytic progenitors (Fig. 2e, Extended Data Fig. 4a).

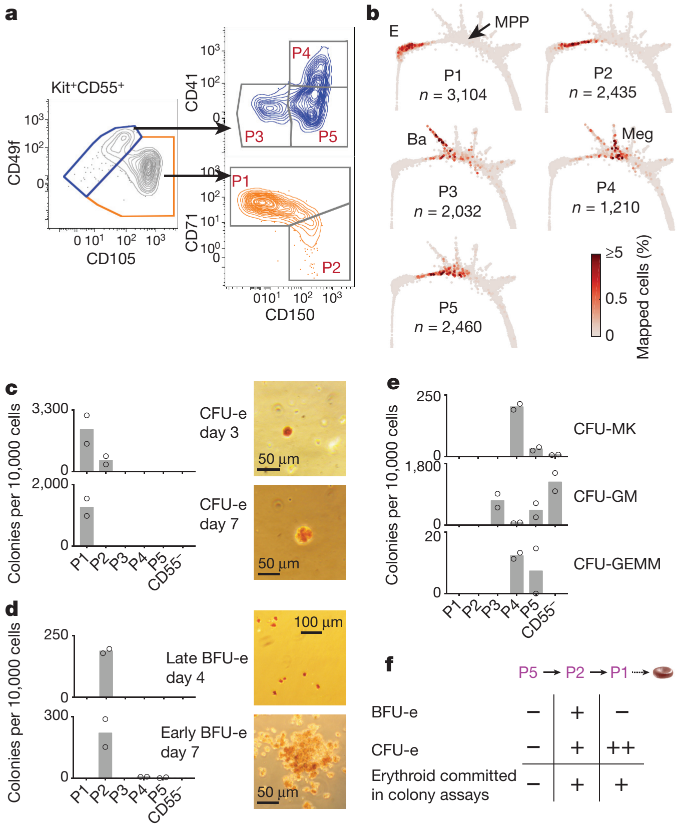


Figure 2 | A novel sorting scheme isolates erythroid progenitors. **a**, Kit$^+$CD55$^+$ bone marrow cells were sorted into gates P1–P5, and profiled using scRNA-seq. **b**, P1–P5 single-cell transcriptomes localized to their most similar counterparts on the SPRING graph. **c–e**, Colony formation by unifocal-erythroid (**c**), multifocal-erythroid (**d**) and non-erythroid (**e**) P1–P5 and Kit$^+$CD55$^-$ cells. Bars represent the mean of two independent experiments (individual circles), each performed in triplicate. Images show representative erythroid colonies, stained for haemoglobin with diaminobenzidine. Colonies: CFU-MK, megakaryocytic (Extended Data Fig. 4a); CFU-GM, granulocytic/monocytic; CFU-GEMM, mixed myeloid. **f**, Summary of erythroid colony potential of FACS subsets.

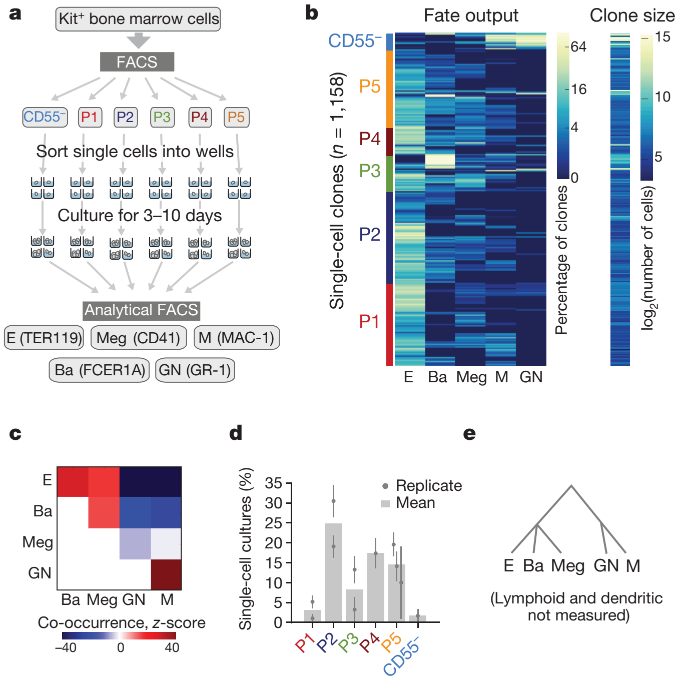


Figure 3 | Predicted fate couplings confirmed by single-cell fate assays. **a**, Schematic of single-cell liquid cultures, measuring clonal output with the indicated antibodies. **b**, Lineage output (left) and size (cell number, right) of each clone (rows). **c**, Concurrently occurring fates in single-cell culture, computed by comparing the number of clones from **b** that produce a pair of fates to the number expected following randomization. **d**, Fraction of bipotent erythroid–Ba/Mast clones in P1–P5, containing erythroid and either basophil or mast cells, but no other fates. Individual points and error bars show the expected value and s.e.m. from independent single-cell sorting experiments. Bars represent the mean of two (P1, P2, P3) or three (P5) independent experiments; a single experiment was performed for P4 and CD55$^-$. **e**, A cell-state hierarchy that is based on the concurrent occurrence of fates in single-cell cultures.

To test HPC fate potential further, we sorted single Kit$^+$ cells into liquid culture wells in the presence of cytokines that support myeloid and erythroid differentiation (Fig. 3a). We assayed the clonal output of 1,158 single cells by FACS (Fig. 3b, Extended Data Fig. 4b). Unipotential clones for the erythroid, Ba/Mast, megakaryocytic, and granulocytic neutrophil and monocytic (GN/M) lineages largely originated in the P1/P2, P3, P4 and CD55$^-$ subpopulations, respectively, consistent with predictions (Figs 1c, 2b, 3b). Many clones contained multiple lineages, with strong, statistically significant couplings between the erythroid, Ba/Mast and megakaryocytic cell fates on the one hand, and the GN/M fates on the other (Fig. 3c; absolute *z*-score value of more than 10 when compared to randomized data), consistent with both known (erythroid and megakaryocytic, GN/M) and novel (erythroid and Ba/Mast) PBA predictions (Fig. 1c, e, f). Progenitors with erythroid and Ba/Mast output were enriched in the P2 and P5 subpopulations (Fig. 3d), which map close to the erythroid and Ba/Mast branch point in the scRNA-seq data (Figs 1c, 2b), and were depleted in the CD55$^-$ population, as predicted. We found similar results in bulk liquid cultures (Extended Data Fig. 4c). Notably, the new Ba/Mast differentiation pathway suggested by our data does not rule out Ba/Mast formation by the traditional route, as some clones gave rise to both granulocytic neutrophil and Ba/Mast lineages. These results suggest that erythroid, Ba/Mast and megakaryocytic fates are coupled transcriptionally and functionally, while being anti-coupled to the GN/M fates, and that scRNA-seq data can be used to generate successful predictions of HPC states and fates.

The erythroid differentiation trajectory

Integrating the cell fate assays and scRNA-seq analysis, we partitioned the continuum of cell states between MPPs and ETD into three stages (Fig. 4a): (1) erythroid–basophil–megakaryocyte-biased progenitors (EBMPs), (2) early erythroid progenitors (EEPs) and (3) committed erythroid progenitors (CEPs). EBMPs are oligopotent cells near the branch points of the megakaryocytic and basophil lineages, which are biased away from the GN/M fates and strongly represented in the P5 and P2 subpopulations (Fig. 2b). EEPs occupy a narrow region of the graph, just past the final non-erythroid fate branch point; they form most of the P2 subpopulation and are functionally BFU-e (Fig. 2b–d). CEPs constitute the majority of unipotential erythroid progenitors, form most of the P1 subpopulation, and are functionally CFU-e (Fig. 2b–d).

To establish the transcriptional events of the erythroid trajectory, we created a smoothed time series for every gene from MPP to ETD, akin to published pseudotemporal-ordering algorithms[35–37] (Fig. 4b). Known erythroid regulators recapitulated the expected expression dynamics (Fig. 4c). *Gata1* and the erythropoietin (EPO) receptor, *Epor*, were induced early, concurrent with suppression of *Spi1* (which encodes PU.1) and *Gata2* (ref. 38). The transition to ETD was marked by a sharp induction of erythroid genes such as α-globin (*Hba-a1*). We validated the expression of canonical transcription factors in sorted P1–P5 subpopulations, including the early expression of *Gata1* (Extended Data Fig. 5a, b). We further established that a graded increase in *Tfrc* (which encodes CD71) is a reliable marker of continuous progression through the EEP and CEP stages; transcriptomes of sorted CD71$^{\mathrm{high}}$ P1 cells map to late CEP stage and CD71 gradually increases in sorted P2 and P1 cells differentiating *in vitro* (Extended Data Fig. 5c, d).

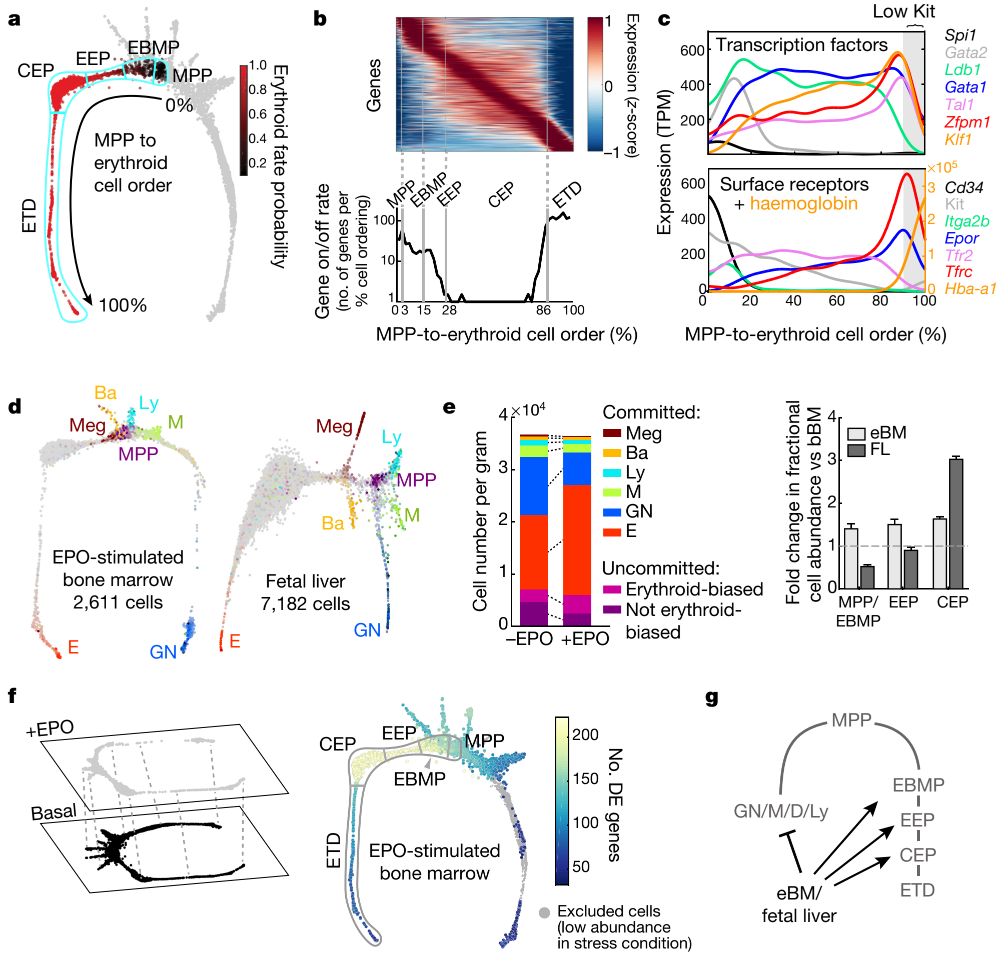


Figure 4 | Stages of early erythropoiesis and the global erythroid stress response. a, Stages of the erythroid trajectory between MPPs and ETD. EBMPs, erythroid–basophil–megakaryocyte-biased progenitors. The SPRING plot shows PBA-predicted erythroid fate probability. **b**, Top, dynamically varying genes (rows), ordered by peak expression, in cells (columns) ordered from MPP to ETD. Gene expression data were smoothed using a Gaussian kernel. Bottom, the number of genes turning on or off (density of expression inflection points) throughout the progression from MPP to ETD. The *x* axis represents PBA-predicted differentiation ordering of cell transcriptomes, uniformly spaced from the least (0%) to the most (100%) differentiated. **c**, Gene expression traces for established erythroid genes. **d**, SPRING plots of EPO-stimulated bone marrow (eBM) and fetal liver samples. Cells coloured as in Fig. 1b. **e**, Left, erythroid lineage expansion at the expense of non-erythroid cells (see Extended Data Fig. 7). Among uncommitted cells, erythroid-biased progenitors increased, whereas the remainder diminished. Committed cells were defined by a PBA-predicted erythroid fate probability greater than 0.5. Right, change in abundance of each progenitor stage relative to basal bone marrow. Error bars are the sampling s.e.m. (one sample per condition). **f**, EPO-stimulated differential gene expression. Cells from EPO-stimulated bone marrow were first mapped onto the basal bone marrow SPRING plot, and then differentially expressed (DE) genes were analysed. **g**, Summary of the stress erythropoiesis response.

A further, sharp induction of *Tfrc* takes place at the transition to ETD (Fig. 4c).

Of the approximately 4,500 genes that varied significantly along the erythroid trajectory (Supplementary Table 3), a large group was induced at the onset of the CEP stage, and sharply suppressed at the CEP to ETD transition (Fig. 4b). It contained the most dominant dynamic gene clusters and was enriched for cell cycle and growth-related genes, including those involved in mTOR signalling, nucleotide metabolism and DNA replication (Extended Data Figs 5e, 6a, b, Supplementary Table 4). These pathways suggest that CEPs, which are the most abundant cells in early erythropoiesis, act as an ‘amplification’ module. Our analysis predicts new epigenetic and transcriptional regulators of the erythroid lineage (Extended Data Fig. 6, Supplementary Table 4), and shows that although GATA1 is expressed early in the erythroid trajectory, most of its canonical targets are induced only at the transition to ETD. Taken together, the temporal ordering of the single-cell transcriptomes recapitulates known events of early erythropoiesis and uncovers a dedicated CEP transcriptional program that is distinct from the ETD program.

Erythroid stress generates a broad response

Using scRNA-seq, we examined two model systems of accelerated, or stress, erythropoiesis: the mid-gestation fetal liver ($n = 7{,}182$ cells post-filter), in which erythropoiesis is a rate-limiting factor of fetal growth; and bone marrow from mice treated with EPO for 48 h, stimulating red blood cell production ($n = 2{,}611$ cells post-filter). SPRING graphs revealed a remarkable conservation of the key features of the haematopoietic hierarchy and erythroid differentiation during stress (Fig. 4d). The proportion of erythroid-trajectory cells increased with stress (Fig. 4d, e). In the fetal liver, the increase was predominantly in CEPs, whereas in EPO-stimulated bone marrow, all erythroid-trajectory cells increased in abundance, including uncommitted MPPs and EBMPs. We found that the absolute number of Kit^+ cells in EPO-stimulated bone marrow did not change, indicating that the increase in erythroid-trajectory cells came at the expense of other cell lineages (Fig. 4e, Extended Data Fig. 7). A number of mechanisms could account for this, including altered intrinsic fate bias of MPPs[39,40].

EPO addition altered gene expression principally in EEPs and CEPs, but also in EBMPs and MPPs (Fig. 4f), in which targets of

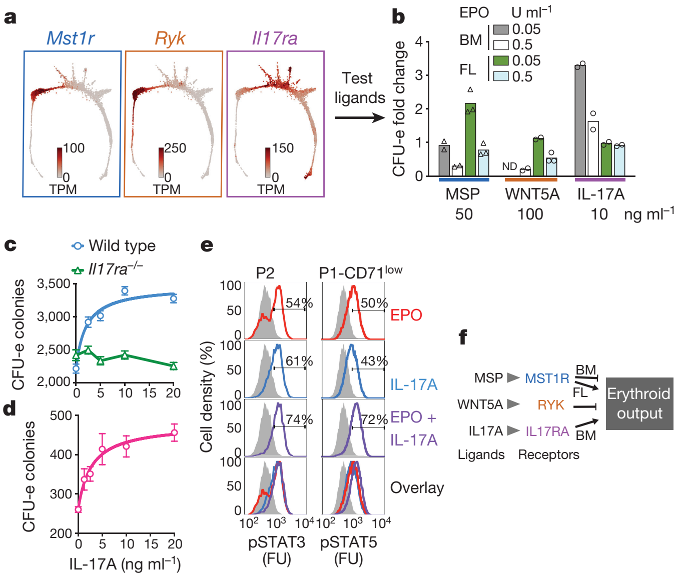


Figure 5 | Novel growth factor regulators of early erythropoiesis. **a**, Expression patterns for *Mst1r*, *Ryk* and *Il17ra* (see Extended Data Fig. 9a, b). TPM, transcripts per million. **b**, Effect of MSP, WNT5A or IL-17A on EPO-dependent CFU-e colony formation. Bars represent the mean of two or three independent experiments (individual data points), each performed in quadruplicate (for full analysis see Extended Data Fig. 9c). **c**, The IL-17A response is lost in *Il17ra*$^{-/-}$ bone marrow. Data are mean ± s.d. per 500,000 bone marrow cells plated in triplicate in the presence of EPO (0.05 U ml^{-1}) and are representative of two independent experiments. **d**, IL-17A stimulates CFU-e formation in freshly isolated human bone marrow mononuclear cells. Data are mean ± s.d. per 85,000 cells plated in triplicate. **e**, IL-17A-mediated phosphorylation of STAT3 and STAT5 (pSTAT3 and pSTAT5). Fresh bone marrow cells were starved of cytokines for 3 h, and then stimulated with EPO, IL-17A or both; FACS profiles are for baseline (starved, shown in grey), and 60 min after stimulation (in colour). Profiles are representative of two independent experiments, each performed in duplicate. FU, fluorescence units. **f**, Summary of growth factor effects on erythroid output.

CCAAT-enhancer-binding protein β (C/EBPβ), a transcription factor that biases differentiation away from erythroid and megakaryocytic fates[41], were downregulated. We identified both known[42,43] and new stress-responsive genes, together with their precise localization within the erythroid trajectory (Extended Data Fig. 8, Supplementary Table 5).

Taken together, we found that the cell-state branching structure is maintained during accelerated erythropoiesis. In MPPs and throughout the ensuing erythroid progression, we identified changes in gene expression and in cell abundance in response to EPO well beyond the currently known mechanism of EPO-driven erythropoietic expansion[42,44].

Growth factor regulators of early erythropoiesis

We screened EEPs and CEPs for gene expression of cell-surface receptors with known ligands using qRT–PCR, identifying three such receptors encoded by *Ryk*, *Mst1r* and *Il17ra* (Fig. 5a, Extended Data Fig. 9a, b). RYK and MST1R have previously been reported in CFU-e, but their function remains unknown[45,46]. However, the expression of an IL-17A receptor by EEPs has not, to our knowledge, been documented. We stimulated RYK, MST1R and IL-17RA with their respective ligands, WNT5A, macrophage-stimulating protein (MSP) and IL-17A, using erythroid colony formation as a readout (Fig. 5b, Extended Data Fig. 9c). In the fetal liver sample, in the presence of low levels of EPO (50 mU ml^{-1}), MSP doubled the number of CFU-e colonies, equivalent to a tenfold increase in EPO concentration. MSP was inhibitory in other contexts, and WNT5A was a potent inhibitor of erythroid colony formation in both the fetal liver and bone marrow samples. By contrast, IL-17A mediated a marked potentiation of adult bone marrow CFU-e colony formation, quadrupling colonies at low levels of EPO (50 mU ml^{-1}), and increasing them by approximately 50% at high levels of EPO (500 mU ml^{-1}).

The stimulatory effect of IL-17A required endogenous IL-17RA (Fig. 5c) and was also evident in human bone marrow (Fig. 5d). Furthermore, IL-17A stimulation was saturable, with a low half-maximum effective concentration (EC$_{50}$) (60 pM), consistent with high-affinity binding of IL-17A to IL-17RA. IL-17A induced rapid phosphorylation of the intracellular signalling mediators STAT3 and STAT5 in CEPs and EEPs (Fig. 5e), and western blotting of freshly sorted CEP P1 and EEP P2 cells revealed expression of IL-17RA (Extended Data Fig. 9d). Taken together, our findings suggest previously unknown regulation of EEPs and CEPs through the expression of a number of growth factor receptors.

Cell cycle remodelling during erythroid development

In a final analysis, we asked what governs progression through the CEP stage and its termination during ETD. We previously reported that the onset of ETD in the fetal liver occurs within a single S phase, and is dependent on S-phase progression[47]; furthermore, this unique S phase is shorter and faster than the S phase in pre-ETD cells[48,49]. These conclusions, which are based on the analysis of large fetal liver subpopulations, predict that CEP exit should show an S phase signature. In our scRNA-seq data, we found that the expression levels of genes that mark the G1/S, S, G2 and G2/M cell cycle phases form a sequence of close, sharp peaks during CEP exit, probably representing a single cell cycle (Fig. 6a, b). This and the following results hold even when cell cycle genes are omitted for ordering the erythroid trajectory (Extended Data Fig. 10a–c). Notably, by reversibly inhibiting DNA replication, we found that the CEP-to-ETD transition in adult bone marrow not only synchronized with, but also depended on, S phase progression (Extended Data Fig. 10d–f).

The scRNA-seq data revealed that changes to cell cycle machinery occur throughout the CEP stage, perhaps in preparation for the switch to ETD. Genes with expression levels that most closely correlate with CEP progression (Supplementary Table 6) are enriched for Gene Ontology terms associated with DNA replication. Notably, regulators of S phase and the G1/S-phase transition increase steadily through the CEP stage, including cyclin E1 (encoded by *Ccne1*), cyclin A2 (*Ccna2*) and MCM helicase subunits (*Mcm2*–*Mcm7*). Conversely, regulators of the G1 phase such as cyclin D2 (*Ccnd2*) and cyclin-dependent kinase 6 (*Cdk6*) decrease steadily (Fig. 6c). To investigate these findings, we labelled S-phase cells *in vivo* with the nucleotide analogue BrdU, and analysed the cell cycle distribution of cells as they progressed through the EEP and CEP stages (Fig. 6d–f). We found a graded but notable increase in the fraction of cells in S phase, whereas the number of G1 cells correspondingly decreased. Results were similar in both EPO-stimulated bone marrow and fetal liver samples (Extended Data Fig. 10g). There was no significant change in the length or speed of S phases, as evidenced by stable intra-S-phase levels of BrdU[48] (Fig. 6f), suggesting that cells spend more time in S phase as a result of G1 shortening. Western blotting of sorted P1 and P2 fractions confirmed that the expression of key S-phase regulators increased with developmental progression in EEPs and CEPs (Extended Data Fig. 10h). Taken together, our data suggest that progression through the erythroid trajectory is associated with extensive remodelling of the cell cycle (Fig. 6g).

Discussion

Our scRNA-seq analysis reveals that HPCs occupy a continuum of transcriptional cell states, branching towards seven fates. Certain cell fate potentials are correlated, supporting a hierarchical view of haematopoiesis, with MPPs diverging either towards myeloid and

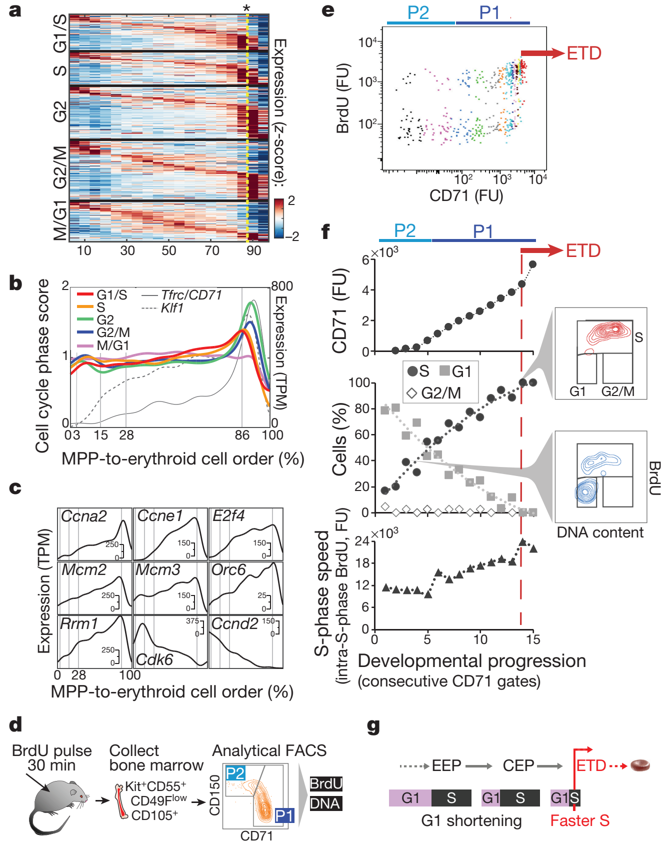


Figure 6 | Extensive remodelling of the cell cycle during erythroid development. **a**, Cell cycle phase-specific genes[52], ordered by peak expression, reveal cell cycle synchronization with the CEP to ETD transition (indicated by an asterisk). **b**, Mean expression of all genes specific to each cell cycle phase (as in **a**), traced along the erythroid trajectory. The transition to ETD is marked by a sharp induction of *Tfrc*. **c**, Representative cell cycle genes that are correlated or anti-correlated with progression along the erythroid trajectory. **d**, Schematic for cell cycle analysis of erythroid progenitors *in vivo*. Bone marrow was collected and fixed 30 min after BrdU injection; P1 and P2 cells were analysed for BrdU incorporation and DNA content. **e**, BrdU-labelled S-phase cells, as in **d**. Cell colouring represents consecutive 7-percentile gates of increasing CD71, reflecting progression through the EEP (P2) and CEP (P1) stages (Extended Data Fig. 5c, d). The transition to ETD (red arrow) is marked by a sharp increase in CD71, and synchronization in the S phase (BrdU^{+}). **f**, CD71 expression (top), cell cycle phase distribution (middle), and intra-S-phase DNA synthesis rate (bottom), for all gates in **e**. Insets show representative FACS plots of cell cycle distribution. Data are representative of three independent experiments. For similar analyses of EPO-stimulated bone marrow and fetal liver samples, see Extended Data Fig. 10g. **g**, Summary of cell cycle remodelling during early erythropoiesis and the S-phase-dependent switch to ETD.

lymphoid fates, or towards the erythroid, megakaryocyte and Ba/Mast cell fates. Yet unlike the classical models of haematopoiesis, HPCs do not separate into discrete and homogenous stages. The coupling of specific cell fates, which we validated with single-cell fate assays, is a critical feature by which our model differs from recent models of haematopoiesis, in which unilineage progenitors arise directly from MPPs. Our model also explains historical hierarchical interpretations of haematopoiesis, which were based on fate assays of FACS-gated populations, averaging the fate couplings of their constituent progenitors. Of note, the continuous nature of the scRNA-seq data does not rule out the existence of discrete epigenetic or signalling states among HPCs, if their lifetime in single cells is comparable to, or shorter than, the lifetime of mRNA molecules (in the range of hours to approximately 1 day).

We delineated the continuous differentiation trajectory of the erythroid lineage, from its origins in MPPs, through EBMPs, to unipotential EEPs and CEPs, which we show correspond to the unipotential BFU-e and CFU-e, respectively. The dominant CEP stage expresses a distinct transcriptional program and is probably a regulator of erythroid output, as evidenced by both its expansion under stress, and by novel growth factor receptors that regulate CFU-e numbers. In particular, our finding of strong stimulation by the pro-inflammatory IL-17RA contributes to the growing evidence of a complex interplay between erythropoiesis and inflammation[50,51]. We further identified the cell cycle as a key process in both the progression and termination of the CEP stage. Developing CEPs spend an increasing fraction of their time in S phase, as a result of G1 shortening; their transition to ETD in an abrupt transcriptional switch is dependent on a single, short S phase. We speculate that the cell cycle may set the context for activation of transcription factors that are induced earlier in the erythroid trajectory. Taken together, our single-cell approach allowed us to make detailed predictions that we validated to reveal novel fundamentals of early haematopoietic differentiation, as well as practical methods for further isolation and study of these cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 October 2016; accepted 11 January 2018.

Published online 21 February 2018.

1. Fujiwara, Y., Browne, C. P., Cunniff, K., Goff, S. C. & Orkin, S. H. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc. Natl Acad. Sci. USA* **93,** 12355–12358 (1996).
2. Liu, Y. *et al.* Suppression of Fas–FasL coexpression by erythropoietin mediates erythroblast expansion during the erythropoietic stress response *in vivo*. *Blood* **108,** 123–133 (2006).
3. Chen, K. *et al.* Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc. Natl Acad. Sci. USA* **106,** 17413–17418 (2009).
4. Hara, H. & Ogawa, M. Erythropoietic precursors in mice under erythropoietic stimulation and suppression. *Exp. Hematol.* **5,** 141–148 (1977).
5. Gregory, C. J., McCulloch, E. A. & Till, J. E. The cellular basis for the defect in haemopoiesis in flexed-tailed mice. III. Restriction of the defect to erythropoietic progenitors capable of transient colony formation *in vivo*. *Br. J. Haematol.* **30,** 401–410 (1975).
6. Pronk, C. J. *et al.* Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell* **1,** 428–442 (2007).
7. Flygare, J., Rayon Estrada, V., Shin, C., Gupta, S. & Lodish, H. F. HIF1α synergizes with glucocorticoids to promote BFU-E progenitor self-renewal. *Blood* **117,** 3435–3444 (2011).
8. Li, J. *et al.* Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. *Blood* **124,** 3636–3645 (2014).
9. Mori, Y., Chen, J. Y., Pluvinage, J. V., Seita, J. & Weissman, I. L. Prospective isolation of human erythroid lineage-committed progenitors. *Proc. Natl Acad. Sci. USA* **112,** 9638–9643 (2015).
10. Guo, G. *et al.* Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* **13,** 492–505 (2013).
11. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).
12. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163,** 1663–1677 (2015).
13. Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*. *Nature* **518,** 542–546 (2015).
14. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351,** aab2116 (2016).
15. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128,** e20–e31 (2016).
16. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19,** 271–281 (2017).
17. Mercier, F. E. & Scadden, D. T. Not all created equal: lineage hard-wiring in the production of blood. *Cell* **163,** 1568–1570 (2015).
18. Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91,** 661–672 (1997).
19. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404,** 193–197 (2000).
20. Adolfsson, J. *et al.* Identification of Flt3^{+} lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121,** 295–306 (2005).

21. Huang, W., Cao, X., Biase, F. H., Yu, P. & Zhong, S. Time-variant clustering model for understanding cell fate decisions. *Proc. Natl Acad. Sci. USA* **111,** E4797–E4806 (2014).
22. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl Acad. Sci. USA* **111,** E5643–E5650 (2014).
23. Shin, J. *et al.* Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17,** 360–372 (2015).
24. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*. **44,** e117 (2016).
25. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* **17,** 106 (2016).
26. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13,** 845–848 (2016).
27. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33,** 269–276 (2015).
28. Khoramian Tusi, B. & Socolovsky, M. High throughput single-cell fate potential assay of murine hematopoietic progenitors *in vitro*. *Ex. Hematol*. https://doi.org/10.1016/j.exphem.2018.01.005 (2018)
29. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
30. Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1,** 661–673 (1994).
31. Papayannopoulou, T., Brice, M., Broudy, V. C. & Zsebo, K. M. Isolation of c-kit receptor-expressing cells from bone marrow, peripheral blood, and fetal liver: functional properties and composite antigenic profile. *Blood* **78,** 1403–1412 (1991).
32. Weinreb, C., Wolock, S. & Klein, A. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* (2017).
33. Weinreb, C., Wolock, S., Khoramian Tusi, B., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single cell snapshots. *Proc. Natl Acad. Sci. USA.* http://doi.org/10.1073/pnas.1714723115 (2018).
34. Yanez, A. *et al.* Granulocyte-monocyte progenitors and monocyte-dendritic cell progenitors independently produce functionally distinct monocytes. *Immunity* **47,** 890–902.e4 (2017).
35. Magwene, P. M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19,** 842–850 (2003).
36. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–725 (2014).
37. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
38. Bresnick, E. H., Lee, H.-Y., Fujiwara, T., Johnson, K. D. & Keles, S. GATA switches as developmental drivers. *J. Biol. Chem.* **285,** 31087–31093 (2010).
39. Li, P. *et al.* Regulation of bone marrow hematopoietic stem cell is involved in high-altitude erythrocytosis. *Exp. Hematol.* **39,** 37–46 (2011).
40. Grover, A. *et al.* Erythropoietin guides multipotent hematopoietic progenitor cells toward an erythroid fate. *J. Exp. Med.* **211,** 181–188 (2014).
41. Mancini, E. *et al.* FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *EMBO J.* **31,** 351–365 (2012).
42. Koulnis, M., Porpiglia, E., Hidalgo, D. & Socolovsky, M. in *A Systems Biology Approach to Blood,* Vol. 844 (eds Corey, S. J.*et al.*) Ch. 3, 37–58 (Springer New York, 2014).
43. Agosti, V., Karur, V., Sathyanarayana, P., Besmer, P. & Wojchowski, D. M. A KIT juxtamembrane PY567-directed pathway provides nonredundant signals for erythroid progenitor cell development and stress erythropoiesis. *Exp. Hematol.* **37,** 159–171 (2009).
44. Koury, M. J. & Bondurant, M. C. Erythropoietin retards DNA breakdown and prevents programmed death in erythroid progenitor cells. *Science* **248,** 378–381 (1990).
45. Yee, K., Bishop, T. R., Mather, C. & Zon, L. I. Isolation of a novel receptor tyrosine kinase cDNA expressed by developing erythroid progenitors. *Blood* **82,** 1335–1343 (1993).
46. van den Akker, E. *et al.* Tyrosine kinase receptor RON functions downstream of the erythropoietin receptor to induce expansion of erythroid progenitors. *Blood* **103,** 4457–4465 (2004).
47. Pop, R. *et al.* A key commitment step in erythropoiesis is synchronized with the cell cycle clock through mutual inhibition between PU.1 and S-phase progression. *PLoS Biol*. **8,** e1000484 (2010).
48. Hwang, Y. *et al.* Global increase in replication fork speed during a p57KIP2-regulated erythroid cell fate switch. *Sci. Adv*. **3,** e1700298 (2017).
49. Shearstone, J. R. *et al.* Global DNA demethylation during mouse erythropoiesis *in vivo*. *Science* **334,** 799–802 (2011).
50. Nemeth, E. & Ganz, T. Anemia of inflammation. *Hematol. Oncol. Clin. North Am.* **28,** 671–681 (2014).
51. Liang, R. *et al.* A systems approach identifies essential FOXO3 functions at key steps of terminal erythropoiesis. *PLoS Genet.* **11,** e1005526 (2015).
52. Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13,** 1977–2000 (2002).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by a Leukemia and Lymphoma Society Scholar award (1728-13) and R01DK100915 and R01099281 (M.S.). A.M.K. is supported by a BW Fund CASI award and an Edward J Mallinckrodt Foundation Grant. S.L.W. and C.W. are supported by National Institutes of Health (NIH) training grant 5T32GM080177-07.

Author Contributions M.S. and A.M.K. designed the experiments and supervised the project. B.K.T., S.L.W., Y.H., D.H. and R.Z. performed experiments including inDrops (B.K.T., R.Z., S.L.W.), FACS and antibody screening (B.K.T., D.H.), single-cell fate assays and cell cycle analysis (B.K.T.), western blotting (Y.H.), qRT–PCR (B.T.K.), pSTAT3/5 (Y.H., D.H.) and colony assays for novel growth factors (Y.H.). S.L.W. and C.W. performed single-cell data analysis, informatics and PBA modelling. A.W. and J.R.H. provided *Il17ra*$^{-/-}$ mice. B.K.T., S.L.W., C.W., Y.H., D.H., A.M.K. and M.S. prepared figures and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.M.K. (allon_klein@hms.harvard.edu) or M.S. (merav.socolovsky@umassmed.edu).

Reviewer Information *Nature* thanks B. Göttgens, F. Hamey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size, the experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Ethical compliance. All mouse experiments described in this project fully comply with the mouse protocol issued to the Socolovsky laboratory by the Institutional Animal Care and Use Committee (IACUC) of the University of Massachusetts Medical School.

Mice for scRNA-seq. For the basal bone marrow (bBM) sample, and for the sorted P1–P5 populations, bone marrow was collected from 8-week-old adult BALB/cJ female mice (Jackson Laboratories). For the EPO-stimulated bone marrow (eBM) sample, 8-week-old adult BALB/cJ female mice were injected with EPO (Procrit, Amgen) subcutaneously once every 24 h for a total of 48 h, at 100 U per 25 g body weight. For the fetal liver (FL) sample, BALB/cJ female mice were set up for timed pregnancies, and fetal livers were collected on embryonic day 13.5.

Cell preparation for scRNA-seq. *Tissue collection.* For bone marrow preparation, femurs and tibiae were collected immediately following euthanasia, and placed in cold (4 °C) 'staining buffer' (PBS containing 0.2% bovine serum albumin (BSA) and 0.08% glucose). Bones were flushed using a 2-ml syringe with a 26-gauge needle and then crushed with a pestle and mortar to obtain all cells. Collected bone marrow cells were filtered through a 40-μm strainer and washed in cold 'Easy Sep' buffer (PBS; 2% fetal bovine serum (FBS); 1 mM EDTA). Fetal livers were prepared by mechanical dissociation in staining buffer and a washing in 'Easy Sep' buffer.

Positive selection for Kit$^+$ cells. Bone marrow and fetal liver cell samples were each enriched for Kit-expressing cells using magnetic beads, with the Mouse Biotin Selection Kit (STEMCELL Technologies, 18556) and Biotin Rat Anti-Mouse CD117 Antibody (clone 2B8, BD Bioscience), following the manufacturer's protocol.

Density gradient centrifugation. Following magnetic bead selection, dead cells and debris were removed from the bone marrow and fetal liver samples using density centrifugation in OptiPrep (Sigma, D1556). In brief, cells were re-suspended in 0.5 ml staining buffer, mixed with 1 ml of 40% of OptiPrep in PBS, and placed in a 5-ml tube. The cell suspension was carefully over-layered with 2 ml of 20% OptiPrep solution, and 1 ml of 5% OptiPrep solution, and centrifuged at 800*g* for 15 min (centrifuge break off). The top visible cell band that formed during centrifugation contained the live, Kit$^+$ single cells (confirmed by flow cytometric analysis). This layer was carefully aspirated and used directly in the inDrops[29] platform.

Single-cell transcriptome droplet microfluidic barcoding using inDrops. For scRNA-seq, we used inDrops[29] following a previously described protocol[53] with the modifications summarized in Supplementary Table 7. Following droplet barcoding reverse transcription, emulsions were split into aliquots of approximately 1,000 single-cell transcriptomes and frozen at −80 °C. Two batches of Kit$^+$ libraries were prepared, referred to as batch 1 (bBM, $n=840$ cells; eBM, $n=1,141$ cells; FL, $n=1,953$ cells) and batch 2 (bBM, $n=4,592$ cells; eBM, $n=1,314$ cells; FL, $n=7,529$ cells) in Supplementary Table 7. These cell numbers correspond to the final number of transcriptomes detected upon sequencing (see 'Cell filtering and data normalization'), and were in agreement with estimated inputs.

For the FACS subsets P1, P1-CD71$^{\text{high}}$, P2, P3, P4, and P5 (referred to collectively as 'P1–P5'), all libraries were prepared in parallel, with a total of 16,206 cell barcodes detected in the sequencing data before filtering (P1, $n=5,733$ cells; P1-CD71$^{\text{high}}$, $n=1,631$ cells; P2, $n=2,630$ cells; P3, $n=2,101$ cells; P4, $n=1,589$ cells; P5, $n=2,522$ cells).

Sequencing and read mapping. The first batch of Kit$^+$ (bBM, eBM and FL) libraries was sequenced on a HiSeq 2000, the remaining Kit$^+$ libraries were sequenced on three NextSeq 500 runs, and all P1–P5 libraries were sequenced on a single NextSeq 500 run. Raw sequencing data (FASTQ files) were processed using the previously described[53] inDrops.py bioinformatics pipeline (available at https://github.com/indrops/indrops), with a few modifications. Bowtie v.1.1.1 was used with parameter -e 100; all ambiguously mapped reads were excluded from analysis and reads were aligned to the Ensemble release 81 mouse mm10 cDNA reference.

Cell filtering and data normalization. Each sample (bBM, eBM, FL and P1–P5) was processed separately. The bBM, eBM and FL samples (referred to collectively as Kit$^+$) were initially filtered to include only abundant barcodes, on the basis of visual inspection of the histograms of total reads per cell (see cell numbers reported in 'Single-cell transcriptome droplet microfluidic barcoding using inDrops'). An additional filtering step removed cells with transcript count totals in the bottom fifth percentile (bBM, $n=271$ cells; eBM, $n=148$; FL, $n=473$). Subsets P1–P5 were filtered only by total transcript counts, with thresholds set by visual inspection of the total counts histograms (see cell numbers reported in 'Single-cell transcriptome droplet microfluidic barcoding using inDrops'). Next, we excluded putatively stressed or dying cells with >10% (bBM, eBM and FL) or >20% (P1–P5) of their transcripts coming from mitochondrial genes (bBM, $n=165$ cells; eBM, $n=45$; FL, $n=698$; P1, $n=2,629$; P1-CD71$^{\text{high}}$, $n=879$; P2, $n=195$; P3, $n=69$; P4, $n=379$; P5, $n=62$).

After cell filtering, we detected the following median number of transcripts and genes per cell, respectively: bBM, 2,989 and 1,539; eBM, 3,082 and 1,552; FL, 8,859 and 2,834; P8, 3,339 and 1,637; P8-CD71$^{\text{high}}$, 4,740 and 2,174; P9, 2,712 and 1,393; P10, 4,641 and 2,158; P11, 1,783 and 1,023; P12, 2,139 and 1,195.

The gene expression counts of each cell were then normalized using a variant of total-count normalization that avoids distortion from very highly expressed genes. Specifically, we calculated $\hat{x}_{i,j}$, the normalized transcript counts for gene j in cell i, from the raw counts $x_{i,j}$ as follows: $\hat{x}_{i,j}=x_{i,j}\overline{X}/X_i$, in which $X_i=\sum_j x_{i,j}$ and $\overline{X}$ is the average of X_i over all cells. To prevent very highly expressed genes (for example, haemoglobin) from correspondingly decreasing the relative expression of other genes, we excluded genes comprising >10% of the total counts of any cell when calculating $\overline{X}$ and X_i.

Exclusion of contaminating cell types and putative cell doublets. To clean up the data for the Kit$^+$ samples, we clustered the single-cell transcriptomes and excluded clusters that were identified as contaminating (non-HPC) cell types and putative cell doublets. No such clusters were detected in the P1–P5 samples. Clustering was performed as follows: we identified the principal variable genes across the entire dataset, as described[29], that is, genes that were highly variable (top 2,000 most variable by *v*-score, a measure of above-Poisson noise (variability)), expressed at non-negligible levels (at least five unique molecular identifier (UMI)-filtered mapped reads (UMIFM) in at least three cells), and which contributed to principal components with eigenvalues greater than those obtained after data randomization ($n=59$, $n=35$ and $n=71$ principal components for bBM, eBM and FL samples, respectively). The expression level for each gene was standardized by a *z*-score transform (mean-subtraction, scaling by s.d.), followed by density-based clustering (DBSCAN)[54,55] on a 2D PCA–tSNE plot (principal component analysis (PCA) followed by *t*-distributed stochastic neighbour embedding (tSNE[56]), as described[29,57]). The tSNE algorithm perplexity parameter was set to 30. Examination of the expression of marker genes in each cluster was then used to identify putative doublets and contaminating cell types.

In the bBM sample, two doublet clusters were identified: one co-expressed markers of mature macrophages and erythrocytes ($n=38$ cells), and the other co-expressed markers of granulocyte and erythroid progenitors ($n=75$ cells). The eBM sample included a cluster of mature macrophages ($n=40$ cells) but no identifiable cluster of doublets. The FL sample contained four contaminating cell types: vascular endothelium, hepatocytes, mesenchymal cells and mature macrophages ($n=769$ cells total), in addition to a small cluster of doublets ($n=18$ cells). Doublets and contaminant cells were excluded from downstream analyses.

To increase confidence that putative doublet clusters were indeed combinations of two single cells, rather than true intermediate or transitional states, we generated simulated 'artificial' doublets by randomly sampling and combining observed transcriptomes.

We then applied PCA–tSNE clustering as described earlier to the union of observed and simulated cells, and identified clusters that were primarily composed of cells with a large number of doublet neighbours (two clusters in bBM, one in FL). These clusters were the same putative doublet clusters identified in the previous paragraph.

Batch correction. Within each Kit$^+$ sample, we observed batch effects between the first and second sequencing runs, with slightly fewer genes detected per cell in the second run compared to the first run. This was consistent with the choice of lower sequencing depth used in the second set of runs, but could also reflect differences in library preparation despite all cells being collected in a single droplet run. To prevent batch effects from distorting subsequent data analysis, for each sample we used the second (larger) batch to select variable genes and to calculate principal component gene loadings. Cells from all batches were then projected into the reduced space, and all subsequent analysis was performed on the reduced principal component space.

Data visualization and construction of *k*-nearest neighbour graphs. After cell filtering, data were prepared for visualization and PBA[33] by constructing a *k*-nearest neighbour (kNN) graph, in which cells correspond to graph nodes and edges connect cells to their nearest neighbours. A kNN graph was constructed separately for each of the three Kit$^+$ samples and for the merged P1–P5 samples (note that the kNN graph for P1–P5 was used only for the visualization in Extended Data Fig. 3).

For the Kit$^+$ samples, genes with mean expression >0.05 and coefficient of variation (s.d./mean) >2 were used to perform PCA down to 60 dimensions (bBM, eBM and FL). For all analyses in this paper, data were *z*-score normalized at the gene level before PCA (qualitatively similar results were also obtained without *z*-score normalization, which weights highly expressed genes more heavily than lowly expressed genes). After PCA, a kNN graph ($k=5$) was constructed by connecting each cell to its five nearest neighbours (using Euclidean distance in the principal component space).

For P1–P5, highly variable genes were filtered using the *v*-score statistic (above-Poisson noise) rather than the coefficient of variation, keeping the top 25% most variable genes and requiring at least three UMIFM to be detected in at least three cells ($n = 3,459$ genes). Additionally, a strong cell cycle signature was observed in the initial graph visualization, manifested by the co-localization of cells expressing G2/M genes (*Ube2c*, *Hmgb2*, *Hmgn2*, *Tuba1b*, *Mki67*, *Ccnb1*, *Top2a*, *Tubb4b*). Therefore, we constructed a G2/M signature score by summing the average *z*-score of these genes, then removed genes that were highly correlated (Pearson $r > 0.2$) with the signature ($n = 31$ genes). Finally, the kNN graph was constructed with $k = 4$ using the first 30 principal components.

The kNN graphs were visualized using a force-directed layout using a custom interactive software interface called SPRING[58]. For the Kit$^+$ samples, several manual steps were taken to improve visualization. It is important to emphasize that the manipulations affect visualization only. All subsequent analyses depend on the graph adjacency matrix, which is not affected by any of the changes to the graph layout. For visualization purposes, we manually extended the length of the megakaryocytic, basophilic and monocytic branches by pinning the position of cells at the end of each branch, and allowing the remaining structure to follow. In the bBM sample, we compressed the CEP 'bulge' region of the graph by bringing its bounding cells together.

Smoothing over the kNN graph. We smoothed data over the kNN graph for gene expression visualization and for one analysis (see 'Global changes in gene expression in stress conditions'). Smoothing was done by diffusing the property of interest (for example, gene expression counts or number of mapped cells) over the graph, as described[59]. In brief, let *A* be the adjacency matrix of the kNN graph, in which $A_{i,j} = 1$ if an edge in the graph connects nodes *i* and *j*. Define A^*as the transition matrix, obtained by row-normalizing *A*:

$$A^*_{i,j} = \frac{A_{i,j}}{\sum_j A_{i,j}}$$

Let E_i be the quantity of interest (for example, expression level) in cell *i*. Then $\boldsymbol{E}^*$, the smoothed vector of *E*, is computed as follows:

$$\boldsymbol{E}^* = \gamma(I - (1-\gamma)A^*)^{-1}E$$

in which γ is a diffusion constant ($\gamma = 0.05$ in all presented analyses) and *I* is the identity matrix.

Formal measure of the continuity of transcriptional states. To demonstrate that the continuous appearance of the Kit$^+$ transcriptomes was not a trivial outcome of our analysis methods, we used the same tools to analyse an scRNA-seq dataset of mature blood cells (peripheral blood mononuclear cells, PBMCs) (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k), which consist of several distinct cell types (Extended Data Fig. 1a). In addition to generating a SPRING plot of the data, we also assessed the interconnectivity of each dataset by examining the behaviour of random walks over the kNN graphs, as previously described[16]. In detail, after subsampling the PBMC data to contain the same number of cells as the bBM dataset, we applied PCA and constructed a kNN graph ($k = 10$) for each dataset. We then simulated 1,000 random walks for each graph and plotted the fraction of nodes (cells) visited as a function of the number of steps (Extended Data Fig. 1a).

PBA. The PBA algorithm calculates a scalar 'potential' for each cell that is analogous to a distance, or pseudotime, from an undifferentiated source, and a vector of fate probabilities that indicate the distance to fate branch points. These fate probabilities and temporal ordering were computed using the Python implementation of PBA (available online from https://github.com/AllonKleinLab/PBA), as described[33].

The inputs to the PBA scripts are a set of comma-separated value (.csv) files encoding: the edge list of a *k*NN graph of the cell transcriptomes (A.csv); a vector assigning a net source/sink rate to each graph node (R.csv); and a lineage-specific binary matrix identifying the subset of graph nodes that reside at the tips of branches (S.csv). These files are provided in the Supplementary Data for the bBM and FL datasets. PBA is then run according to the following steps:

(1) Apply the script 'compute_Linv.py -e A.csv', here inputting edges (flag '-e') from the SPRING kNN graph (see above). This step outputs the random-walk graph Laplacian, Linv.npy.

(2) Apply the script 'compute_potential.py -L Linv.npy -R R.csv', here inputting the inverse graph Laplacian (flag '-L') computed in step (1) and the net source/sink rate to each graph node (flag '-R'). This step yields a potential vector (V.npy) that is used for temporal ordering (cells ordered from high to low potential). The vector R provided in the Supplementary Data was estimated as described in the next section.

(3) Apply the script 'compute_fate_probabilities.py -S S.csv -V V.npy -e A.csv -D 1', here inputting the lineage-specific exit rate matrix (flag '-S'), the potential (flag '-V') computed in step (2), the same edges (flag '-e') used in step (1) and a diffusion constant (flat '-D') of 1. This step yields fate probabilities for each cell.

Figures 1–6 make use of PBA analyses of bBM data. For Fig. 4e and Extended Data Fig. 8, a temporal ordering of erythroid differentiation was generated for the FL dataset using the same steps, with input files that are also provided in Supplementary Data.

Estimation of net source/sink rate vector *R*. A complete definition of the vector $\boldsymbol{R}$ in terms of biophysical quantities has been published previously[33]. In brief, for a gene expression space described by a vector $\boldsymbol{x} = (x_1, x_2,\ldots, x_N)$ giving the expression of each of *N* genes, $\boldsymbol{R}(\boldsymbol{x})$ gives the net imbalance between cell division and cell loss locally for cells with gene expression profile $\boldsymbol{x}$. $\boldsymbol{R}(\boldsymbol{x})$ is corrected for cell enrichment and loss resulting from experimental procedures such as sample enrichment, as follows. In this experiment, all progenitors including HSCs express Kit, but eventually downregulate it as they terminally differentiate. Thus, no cells enter the experimental system other than through proliferation of existing Kit$^+$ HPCs, but the selection for Kit$^+$ cells during sample isolation induces a net sink on cells downregulating Kit expression. For a self-renewing system, cell division and cell loss are precisely balanced, so $\int \boldsymbol{R}(\boldsymbol{x})d\boldsymbol{x} = 0$. To apply PBA, one does not need to estimate $\boldsymbol{R}(\boldsymbol{x})$, but only its value at points x_i at which the *M* cells $i = 1,\ldots, M$ are observed in the scRNA-seq measurement. Thus $\boldsymbol{R}$ is a vector over the cells in the system. For a self-renewing system, the sum over all cells satisfies the same constraint, $\sum_i R_i = 0$.

Estimation of *R*. We assigned negative values to $\boldsymbol{R}$ for the ten cells with the highest expression of marker genes for each of the seven terminal lineages (see Supplementary Table 1 for marker genes), which were separately confirmed to show reduced Kit expression. We assigned different exit rates to each of the seven lineages using a fitting procedure that ensured that cells identified as putative HSCs would have a uniform probability to become each fate. Putative HSCs were identified by the similarity of their transcriptomes to microarray profiles from the ImmGen database (we used SC.LT34F.BM (long-term bone marrow HSCs) for bBM and SC.STSL.FL (short-term FL HSCs) for FL; for more details, see section 'ImmGen Bayesian classifier'). We assigned a single positive value to all remaining cells, with the value chosen to enforce the steady-state condition $\sum_i R_i = 0$. In the fitting procedure, all exit rates are initially set to one and iteratively incremented or decremented until the average fate probabilities of the putative HSCs were within 1% of uniform. The resulting vector $\boldsymbol{R}$ is provided in the Supplementary Data. The separate lineage exit rates were then used to form the lineage-specific exit rate matrix *S*, also provided in the Supplementary Data.

Assignment of PBA fate probabilities and temporal ordering to eBM dataset. For each of the eBM cells we assigned the average temporal order (or potential *V*) and average fate probabilities of the 20 mostly similar bBM cells. To do this, we first carried out a PCA on the bBM cells into 60 dimensions. We then used the gene loadings of the 60 principal components to project the eBM data into the same principal component space. The distance of each eBM cell to each bBM neighbour was then measured by cosine distance in the 60-dimensional sub-space.

ImmGen Bayesian classifier. We used a published microarray profile[60] to search for similar cells in our own dataset using a naive Bayesian classifier, implemented as follows.

The Bayesian classifier assigns cells to microarray profiles on the basis of the likelihood of each microarray profile for each cell, with the likelihood calculated by assuming that individual mRNA molecules in each cell are multinomially sampled with the probability of each gene proportional to the microarray expression value for that gene. Consider a matrix *E* of mRNA counts (UMIs) with *n* rows (for cells) and *g* columns (for genes), and also a matrix *M* with *m* rows (for microarray profiles) and *g* columns (for genes). *M* was quantile normalized and then each microarray profile was normalized to sum to one. Normalization of *E* was performed as described earlier (see 'Cell filtering and data normalization'). The $n \times m$ matrix, S_{ij}, giving the likelihood of each microarray profile *j* for each cell *i* is,

$$S_{ij} = Z_i \prod_{k=1}^{g} M_{jk}^{E_{ik}}$$

where Z_i is a normalization constant that ensures that $\sum_i S_{ij} = 1$.

Computing the haematopoietic lineage tree. We used the fate probabilities from PBA to infer the topology of the haematopoietic lineage tree using an iterative approach (Fig. 1e, f). Each iteration began with a set of fates and a probability distribution over those fates for each cell. For every pair of fates, we computed a fate coupling score (see later) and merged pairs with a score significantly higher than expected under a null model. The merged fates inherited probabilities from the starting fates by simple pairwise addition.

The coupling score between two fates A and B is the number of cells with $P(A)P(B) > \varepsilon$, in which we used a value $\varepsilon = 1/14$ throughout. To generate a null distribution for each fate pair, we computed pairwise coupling scores for 1,000 permutations of the original fate probabilities. The heat maps in Fig. 1e show z-scores with respect to these null distributions.

Analysis of fate-correlated genes at haematopoietic choice points. To discover fate-associated genes at key choice points in haematopoiesis (Extended Data Fig. 2, Supplementary Table 2), we ranked transcription factors and cell-surface markers (transcription factors from http://genome.gsc.riken.jp/TFdb/tf_list.html, cell-surface markers from https://www.thermofisher.com/us/en/home/life-science/cell-analysis/cell-analysis-learning-center/cell-analysis-resource-library/ebioscience-resources/mouse-cd-other-cellular-antigen.html) by their correlation with PBA-predicted fate probability, restricting to cells that were bipotent for the given choice. Specifically, to find transcription factors associated with fate A at an A/B choice point, we first selected cells with $P(A) \times P(B) > \varepsilon$, and then ranked the transcription factors by their correlation with the fate bias ($P(A) - P(B)$). In Supplementary Table 2, we report all genes with Bonferroni-corrected $P < 0.01$ (Pearson correlation coefficient). In Extended Data Fig. 2, we show at most ten genes for any one choice point.

Mapping P1–P5 subsets to the Kit$^+$ graphs. For Fig. 2b, cells from subsets P1–P5 were projected into the same principal component space as the bBM data, then mapped to their most similar Kit$^+$ neighbours. In detail, counts were first converted to TPM for all samples. Then, using only the bBM cells, the 3,000 most variable genes (measured by v-score) with at least three UMIFM in at least three cells were z-score normalized and used to find the top 50 principal components. Next, the P1–P5 subset cells were z-score normalized using the gene expression means and s.d. from the bBM data and transformed into the bBM principal component space. Lastly, each P1–P5 cell was mapped to its closest bBM neighbour in principal component space (Euclidean distance).

Extracting MPP-to-erythroid trajectory cells. To isolate the erythroid trajectory, we defined an MPP-to-erythroid axis in each of the three Kit$^+$ datasets by ordering cells on the basis of their graph distance from unbiased MPPs (cells identified on the basis of the ImmGen classifier as described earlier), and keeping only cells for which the probability of erythroid fate increased or remained constant with graph distance. Graph distance was measured by PBA potential, and starting with the cell closest to the HSC origin, we added the cell with next highest potential to the trajectory if the PBA-predicted erythroid probability for cell i was at least 95% of the average erythroid probability of the cell(s) already in the trajectory.

More formally the procedure is as follows: order all N cells in the experiment from highest to lowest PBA potential V, with decreasing potential corresponding to increasing distance from MPPs[33]. Let E_i be an indicator variable for the membership of ordered cell i in the erythroid trajectory ($E_i = 1$ if cell i is in the trajectory; otherwise, $E_i = 0$). If P_i is the PBA-predicted erythroid probability for ordered cell i, then $E_i = 1$ if

$$P_i > 0.95 \times \frac{\sum_{k<i} P_k \times E_k}{\sum_{k<i} E_k}$$

Cells on the erythroid trajectory were then ordered by decreasing potential. Defining t_j as the index of the jth erythroid-trajectory cell,

$$t_j = 1 + \sum_{k<j} E_k$$

Throughout this Article, we report this cell order (akin to the 'pseudotime' in other publications) as a percentage of ordered cells, with the first, least differentiated cell at 0% and the most mature cell at 100%. This is not meant to suggest that erythroid differentiation ends with this final observed cell.

Identifying dynamically varying genes. For each gene, a sliding window ($n = 100$ cells) across the MPP-to-erythroid ordering was used to identify the windows with maximum and minimum average expression as previously described[57]. A t-test was then performed to assess the statistical significance of the difference in expression levels. To estimate the false discovery rate (FDR), we permuted the order of the cells and repeated the above analysis[57]. For a P value generated by the observed (non-permuted) ordering, the FDR-corrected P value is the fraction of genes from the permuted ordering with that P value or less. Any gene with an FDR-corrected $P < 0.05$ was considered significantly variable.

Identifying stage transitions in the MPP-to-erythroid trajectory. Transition points between stages of erythropoiesis were defined using the frequency of gene inflection points (Fig. 4b), patterns of PBA-predicted fate probabilities (Fig. 1c), and the fate potentials of FACS subsets P1–P5 (Figs 2, 3). However, owing to the continuous nature of the transcriptional states, the locations of these transitions should be considered approximate.

The inflection point density is the number of genes turning on or off at a given point on the trajectory. For each gene, inflection points were identified as the points with maximally increasing or decreasing expression as follows. First, the trajectory of each dynamically varying gene was smoothed using Gaussian smoothing with a width $\sigma = 5\%$ of total trajectory. The gene expression derivative for gene k, denoted x_k, was then computed by taking a ten-cell moving average of the difference between consecutive smoothed gene expression values. Inflection points were then identified as the points with maximum or minimum derivatives for each gene. To exclude maxima or minima resulting from relatively small fluctuations in gene expression, only appreciably large extrema were kept for further analysis. Specifically, the point with the maximum derivative for gene k, $\max(x_k)$, was kept only if

$$\frac{\max(x'_k)}{\text{median}(\text{abs}(x'_k))} > Q$$

Minima were similarly filtered, requiring the ratio to be $< -Q$. We chose a threshold $Q = 6$, but results do not qualitatively change over a range of Q. We then plotted the density of these inflection points over the MPP-to-erythroid axis. Regions with large-scale changes in gene expression have a high density of inflection points, whereas a low density characterizes relatively stable states.

Dynamic gene clustering. Dynamically varying genes were clustered on the basis of their behaviour at the transition points. To prevent overfitting, we used only three transitions (3%, 18% and 86%) by splitting the EEP state and assigning the first and second halves to the EBMP and CEP states, respectively. At each transition, genes were classified as increasing, decreasing or unchanging, giving a total of $3^3 = 27$ possible patterns. After smoothing gene expression traces, the data were binned by calculating the mean expression in each of the four stages. To remove noisy genes or genes that varied little across bins, we calculated the range of binned expression values, $\text{range}(x_{i,\text{binned}}) = \max(x_{i,\text{binned}}) - \min(x_{i,\text{binned}})$, for each gene and proceeded with the top 50% most variable genes. Next, to place all genes on a similar scale, the binned expression values of each gene were divided by the maximum binned value of that gene. Finally, the differences between consecutive bins were thresholded: differences that were greater than 0.15 were called increasing, differences that were less than -0.15 were called decreasing and differences that were between -0.15 and 0.15 were called unchanging.

Gene set enrichment analysis. Each of the 27 gene clusters was used as input for gene set enrichment analysis (GSEA) (hypergeometric test), using all genes as background. Ribosomal genes were excluded from the input, as were predicted genes (gene names starting with 'Gm'). Gene sets from the following lists of the MSigDB v5.1 (ref. 61) dataset were tested for enrichment: Hallmark (h.all.v5.1.symbols.gmt), C2 curated canonical pathways (c2.cp.v5.1.symbols.gmt), C3 transcription factor targets (c3.tft.v5.1.symbols.gmt), and C5 Gene Ontology (c5.all.v5.1.symbols.gmt). Additionally, for the transcription factor target enrichment analysis, we used gene sets from the ChEA database[62].

Cell cycle phase analysis. Genes with periodic expression correlated with the cell cycle in HeLa cells[52] were used to generate a cell cycle phase score for each cell. The list of phase-specific genes was filtered to exclude genes with a mean expression > 25 TPM in cells on the MPP-to-erythroid trajectory. For Fig. 6a, a sliding window average was computed using a window size of 10% MPP-to-erythroid progression (~200 cells) and a jump size of 5%. For Fig. 6b, counts were normalized by the mean expression at the gene level, and smoothed using a Gaussian kernel. Then, a phase score was calculated for each phase (G1/S, S, G2/M, M, M/G1) by averaging the smoothed gene expression traces for the genes specific to that phase.

Testing the influence of cell cycle genes on the MPP-to-erythroid cell order. To test the extent to which cell cycle genes influenced the ordering of cells along the MPP-to-erythroid trajectory, we excluded annotated cell cycle genes as described[63] (cell cycle genes were extracted from the Gene Ontology database (GO:0007049) and Cyclebase[64]), and repeated kNN graph construction and PBA. As shown in Extended Data Fig. 10, the resulting cell order was largely unchanged, as were the dynamics of cell cycle genes.

Identifying genes that change steadily in the CEP stage. To identify genes that are steadily up- or downregulated throughout the CEP (Fig. 6c, Supplementary Table 6), we tested the magnitude of change (slope) and the linearity of change (the error of the actual gene trace from a straight line) for each gene. Restricting analysis to cells in the CEP stage and genes with at least two UMIFM in at least five cells, we fit a linear regression to the ordered gene expression values and also generated a smoothed expression trace using a Gaussian kernel (width $\sigma = 5\%$). We then computed a 'linearity score' for each gene by dividing the slope of the regression line by the root-mean-square error between the regression line and the smoothed trace. Steadily increasing genes receive large positive scores, whereas steadily decreasing genes are assigned large negative scores. Genes that do not change much or that change nonlinearly (for example, sharply increasing only at the end of the stage) receive scores close to 0.

Global changes in gene expression in stress conditions. Cells from eBM and FL (stress samples) were mapped to their most similar bBM counterparts, and differentially expressed genes were identified. Mapping was carried out by applying PCA to the bBM and stress samples and finding the closest 20 bBM neighbours for each stress cell. Specifically, the input genes were the principal variable genes described in the 'Cell filtering and data normalization' section. Count matrices were *z*-score normalized separately for each sample, and PCA was performed on the basal sample to obtain the gene loadings. Using the top 60 principal components, each sample was then transformed using these coefficients, thereby projecting the cells into the same PCA space. To validate this mapping method, we performed the same procedure using different subsets of bBM data as training and test sets (see 'Validation of cross-sample cell mapping' section)

The 20 closest bBM neighbours (Euclidean distance) of each stress cell were found, and for the purpose of comparing gene expression, each of these k (20) neighbours inherited $1/k$ (1/20) of the transcript counts from the mapped stress cell. To enable the comparison of regions of gene expression space (as opposed to comparing single mapped cells to single basal cells), the mapped and original gene expression values were smoothed over the kNN graph, as described in 'Smoothing over the kNN graph'. To avoid comparing gene expression patterns in regions that were relatively unpopulated in the stress sample (for example, parts of the granulocyte branch), we smoothed the number of mapped stress cells per basal cell over the graph and then excluded basal cells with few mapped stress cells (number mapped cells ≤ 9 for eBM and ≤ 20 for FL).

A differential expression score for each cell i and gene j was defined as the maximum-normalized difference between mapped and basal expression, $\hat{x}^*_{i,j}$ and $\hat{x}_{i,j}$, respectively:

$$d_{i,j} = \frac{\hat{x}^*_{i,j} - \hat{x}_{i,j}}{0.5 \times (\max(\hat{x}^*_j) + \max(\hat{x}_j))}$$

A gene level score, D_j, was created by summing over the cells, $D_j = \sum_i d_{i,j}$. Genes were considered differentially expressed if $D_j > \overline{D} + 2 \times \sigma_D$ or $D_j < \overline{D} - 2 \cdot \sigma_D$, in which $\overline{D}$ is the average over all gene level scores D_j and σ_D is the standard deviation.

Then, for each differentially expressed gene, the gene was counted as differentially expressed at a given cell if $d_{i,j} > 0.5 \times \delta_{\text{high}}$ or $d_{i,j} < 0.5 \times \delta_{\text{low}}$, in which δ_{high} is the 99th percentile of D_j and δ_{low} is the 1st percentile of D_j.

Validation of cross-sample cell mapping. To test the accuracy of the method for mapping eBM and FL cells to bBM cells, we divided the bBM sample into a training set (random sample of 75% of the cells) and test set (the remaining 25%). The mapping procedure described in the previous section was then used to map the test set to the training set. As one measure of the accuracy of the mapping, we assigned the test cells the average PBA-predicted fate probabilities and differentiation ordering of the training cells to which they mapped. Both measures were relatively unchanged from their original values (Spearman correlation of 0.97 for the differentiation ordering and >0.95 for each fate probability). As a second measure, we repeated the test for finding global changes in gene expression, using the same gene level score (D_j) cut off as for the eBM sample. This revealed no significantly differentially expressed genes between the training and test sets.

Region-specific differential expression. Before identifying differentially expressed genes, we excluded genes with large batch effects. Although different sequencing depths led to a small change in the average expression of many genes from the first batch to the second, a small number of genes showed major batch effects beyond this, presumably owing to differences in library preparation. We performed a binomial test for differential expression[65] between the two batches of cells and excluded genes with $P < 10^{-50}$, resulting in the removal of 461 genes.

In general, genes can be differentially expressed globally or only in specific cell populations. Particularly when comparing FL to bBM samples, many genes showed global up- or downregulation. To identify differentially expressed genes that are likely to be important specifically for erythropoiesis (or in a particular stage of erythropoiesis), we created a region-specific differential expression score. This score measures the magnitude of the expression difference within a region of interest (ROI) relative to the magnitude outside the region; genes with a larger difference within the ROI than outside of it receive a high score (positive for upregulation, negative for downregulation). For the analyses described here, we tested for differential expression in five ROIs: the stages of the erythroid trajectory, EBMP, EEP, CEP, and ETD and an expanded selection of MPP cells, which included cells with a maximum PBA-predicted lineage probability (for all lineages) less than 0.4, with the exception of cells already included in one of the stages of the erythroid trajectory.

After mapping stress cells to their single closest neighbour in the bBM sample (as described in the previous section), we selected bBM cells in the ROI and the stress cells mapping to them. We first identified genes differentially expressed within the ROI by performing a binomial test for differential expression[65], which tests the probability that a gene is expressed more frequently in one population than another. After correcting for multiple hypothesis testing (Benjamini–Hochberg procedure[66]), we proceeded with genes with an FDR-corrected $P < 0.05$.

To identify genes differentially expressed specifically within the ROI and not elsewhere, we calculated the mean-normalized expression difference for ROI cells and non-ROI cells for the genes found to be significant in the binomial test. For two samples, A (stress) and B (basal), the mean-normalized expression difference of gene i within the ROI, $y_{\text{in},i}$, is

$$y_{\text{in},i} = \frac{\overline{x}^A_{\text{in},i} - \overline{x}^B_{\text{in},i}}{(\overline{x}^A_{\text{all},i} + \overline{x}^B_{\text{all},i})/2}$$

in which $\overline{x}^A_{\text{in},i}$ is the average expression of gene i within the ROI in sample A. A similar score was calculated for cells outside the ROI:

$$y_{\text{out},i} = \frac{\overline{x}^A_{\text{out},i} - \overline{x}^B_{\text{out},i}}{(\overline{x}^A_{\text{all},i} + \overline{x}^B_{\text{all},i})/2}$$

Plotting $y_{\text{in},i}$ against $y_{\text{out},i}$ clearly reveals genes that are more highly differentially expressed within the ROI than without. A single score per gene was computed as follows:

$$\text{score}_i = \begin{cases} \max(y_{\text{in},i} - \max(y_{\text{out},i}, 0), 0)\text{if } y_{\text{in},i} > 0 \\ \min(y_{\text{in},i} - \min(y_{\text{out},i}, 0), 0) \text{ if } y_{\text{in},i} < 0 \end{cases}$$

Intuitively, this score is large and positive if a gene is more strongly upregulated within the ROI than without, large and negative if a gene is more strongly downregulated within the ROI than without and close to 0 otherwise.

To build gene lists for GSEA input, we first selected genes with $\text{score}_i > 0.1 \times \max(\text{score})$ (for upregulated genes) or $\text{score}_i < 0.1 \times \max(\text{score})$ (for downregulated genes) and then used the top 100 genes by binomial test P value.

Flow cytometric sorting for P1–P5 subsets. A detailed protocol of this procedure can be found at the Protocol Exchange[67].

Bone marrow cells from adult BALB/cJ male or female mice (aged 8–12 weeks) were lineage-depleted using the Mouse Streptavidin RapidSpheres Isolation Kit (STEMCELL Technologies 19860A), with the following biotinylated antibodies: anti-CD11B (clone M1/70, BD Biosciences 557395), anti-LY-6G and LY-6C (clone RB6-8C5, BD Biosciences 553125), anti-CD4 (clone RM4-5, BD Biosciences 553045), anti-CD8A (Ly-2) (clone 53-6.7, BD Bioscience 553029), anti-CD19 (clone 1D3, BD Biosciences 553784), anti-TER119 (clone TER119, BD Biosciences 553672).

Lineage-depleted cells were then labelled with the following antibodies in the presence of 1% rat serum: streptavidin Alexa Fluor 488 (Molecular Probes) to mark lineage-positive cells, CD117–APC Cy7 (clone 2B8, Biolegend 105826), TER119–BUV395 (clone TER-119, BD Biosciences 563827), CD71–PE Cy7 (clone RI7217, Biolegend 113812), CD55–AF647 (clone RIKO-3, Biolegend 131806), CD105–PE (clone MJ7/18, Biolegend 120408), CD150–BV650 (clone TC15-12F12.2, Biolegend 115931), CD41–BV605 (clone MWReg30, Biolegend 133921), CD49f (also known as ITGA6)–BV421 (clone GoH3, Biolegend 313624)

After washes, cells were re-suspended in DAPI-containing buffer and sorting was performed using a BD FACSAria II with a 100-μm nozzle. Sorted populations were defined as in Fig. 2a.

qRT–PCR on sorted populations. RNA was prepared from sorted cell subsets using the RNeasy Micro Kit (Qiagen 74004) or TRIzol reagent (Ambion 15596026), and measured with RiboGreen RNA reagent kit (Thermo Fisher Scientific) on the 3300 NanoDrop Fluorospectrometer. cDNA was synthesized using the same amount of input RNA for all samples in a parallel reaction, using the Super Script III first-strand synthesis system for RT–PCR (Invitrogen) with random hexamer primers. The ABI 7300 sequence detection system, TaqMan reagents and TaqMan MGB probes (Applied Biosystems) were used following the manufacturer's instructions. Quantitative PCR was carried on four serial dilutions of each cDNA sample, and the linear part of the template dilution/signal response curve was used to calculate relative mRNA concentrations following normalization to *Actβ*, using the ΔC_t method.

The following TaqMan MGB probes were used: *Mst1r* (Mm00436382_m1), *Ryk* (Mm01238551_m1), *Il17ra* (Mm00434214_m1), *Mt2* (Mm00809556_s1), *Slc26a1* (Mm01198850_m1), *Slc4a1* (Mm00441492_m1), *Trib2* (Mm00454876_m1), *Cd34* (Mm00519283_m1), *Meis1* (Mm00487664_m1), *Hpn* (Mm01152654_m1), *Pf4*

(Mm00451315_g1), *Dntt*, (Mm00493500_m1), *Ms4a2* (Mm00442778_m1), *Elane* (Mm00469310_m1), *S100a9* (Mm00656925_m1), *F13a1* (Mm00472334_m1), *Egr1* (Mm00656724_m1), *Apoe* (Mm01307193_g1), *Ldb1* (Mm00440156_m1), *Zfpm1* (Mm00494336_m1), *Tfrc* (Mm00441941_m1), *Hbb-b1* (Mm01611268_g1), *Alas2* (Mm01260713_m1), *Band3* (also known as *Slc4a1*) (Mm01245920_g1), *Nfe2* (Mm00801891_m1), *Gata1* (Mm01352636_m1), *Gata2* (Mm00492300_m1), *Klf1* (Mm00516096_m1) and *Spi1* (also known as *PU.1*) (Mm00488393_m1).

Colony-formation assays in methylcellulose for P1–P5 and Kit$^+$CD55$^-$ cells. From each freshly sorted cell population, 10,000 cells were mixed with 1 ml MethoCult (M3234, STEMCELL Technologies) supplemented with EPO (2 U ml^{-1}), stem-cell factor (SCF) (50 ng ml^{-1}), IL-3 (10 ng ml^{-1}) and IL-6 (10 ng ml^{-1}). Erythroid (CFU-e or BFU-e) and granulocytic/monocytic colonies were scored from triplicate plates on days three, four and seven of culture. Haemoglobin expression in erythroid colonies was verified by staining with diaminobenzidine *in situ* before scoring.

For megakaryocytes, the colony-formation assay was carried out using MegaCult-C Complete Kit (04970/04972) with added thrombopoietin (TPO) (50 ng ml^{-1}), IL-3 (10 ng ml^{-1}), IL-6 (20 ng ml^{-1}) and IL-11 (50 ng ml^{-1}). From each freshly sorted subset, 10,000 cells were plated in double chamber slides. On day seven of culture, the slides were dehydrated, fixed in ice-cold acetone, and stained for acetylcholinesterase.

Bulk liquid cultures of sorted cell populations. Sorted cells were cultured in IMDM medium in the presence of 20% FCS supplemented with SCF (50 ng ml^{-1}), IL-3 (10 ng ml^{-1}), IL-6 (10 ng ml^{-1}), EPO (2 U ml^{-1}), TPO (50 ng ml^{-1}), IL-11 (50 ng ml^{-1}) and IL-5 (10 ng ml^{-1}) for 7 days. Cells were collected on days two, five and seven, and labelled with the following cell-surface markers for flow cytometric analysis: TER119–BV421 (clone TER-119, Biolegend 116233), CD71–PE Cy7 (clone RI7217, Biolegend 113812), CD117–APC Cy7 (clone 2B8, Biolegend 105826), FCER1A–AF700 (clone MAR-1, Biolegend 134323), CD41–BV605 (clone MWReg30, Biolegend 133921), CD11B–PE Cy5 (clone M1/70, Biolegend 101209), LY 6G/C–FITC (clone RB6-8C5, BD Biosciences 553126).

Single-cell liquid cultures of mouse bone marrow progenitors. Freshly collected mouse bone marrow was labelled with the same antibody scheme as detailed earlier, to allow identification of the Kit$^+$ gates for P1–P5 and CD55$^-$. Single cells were sorted from each of these gates into 96-well plates, retaining index-sorting parameters for each cell, using a BD FACSAria II with a 130-μm nozzle. Cells were cultured for 3–10 days in IMDM with 20% FBS, with the following added growth factors: SCF (50 ng ml^{-1}; recombinant murine SCF, Peprotech 250-03), IL-3 (10 ng ml^{-1}; recombinant murine IL-3, Peprotech 213-13), IL-6 (10 ng ml^{-1}; recombinant murine IL-6, Peprotech 216-16), EPO (2 U ml^{-1}; PROCRIT (epoetin alfa) 606-10-971-8), IL-11 (50 ng ml^{-1}; recombinant murine IL-11, Peprotech 220-11), IL-5 (10 ng ml^{-1}; recombinant murine IL-5, Peprotech 215-15), TPO (50 ng ml^{-1}; recombinant murine TPO, Peprotech 315-14), G-CSF (15 ng ml^{-1}; recombinant murine G-CSF, Peprotech 250-05), GM-CSF (15 ng ml^{-1}; recombinant murine GM-CSF, Peprotech 315-03).

Fresh growth factors were added to the medium of each well on days 4 and 8. The clones in each well were labelled on days 3, 7 or 10, with the same antibody cocktail as described in the 'Bulk liquid cultures of sorted cell populations' section, but with concentrations for each antibody batch that were first optimized with appropriate titrations, to minimize non-specific binding under conditions of low cell number. Clones were analysed using the high throughput sampler (HTS) attachment of the BD LSR II (BD Biosciences).

Fate co-occurrence from single-cell liquid culture data. To measure the statistical significance of fate co-occurrence from the single-cell fate assay data, we used a method similar to that described for calculating fate couplings from the PBA predictions (see 'Computing the haematopoietic lineage tree' section). Because we assayed clonal fate from each FACS subset separately, clones were not represented at the same frequency as in the Kit$^+$ pool (number of clones assayed: CD55$^-$, $n = 58$; P1, $n = 287$; P2, $n = 324$; P3, $n = 125$; P4, $n = 96$; P5, $n = 268$; average frequency in Kit$^+$ population: CD55$^-$ 59.1%, P1 21.4%, P2 6.6%, P3 4.0%, P4 0.8%, P5 4.9%). To adjust for this, we randomly resampled the clone data to ensure clones from each subset were represented in the same proportion as in the Kit$^+$ population (originally: $n = 1{,}158$ clones; after resampling: $n = 8{,}000$ clones). We then computed the observed fate co-occurrence for each fate pair as the number of clones with >2% of cells of the two fates (permitting the presence of other fates as well). Next, we estimated the null distribution by shuffling the data of each fate separately (2,000 replicates) and counting fate co-occurrence as described earlier. Lastly, we calculated the significance of the co-occurrence of each fate pair as the *z*-score of the observed co-occurrence with respect to the null distribution.

In Fig. 3d, the expectation value (E) and s.e.m. for the fraction of bipotent erythroid–basophil cells from each independent experiment were calculated from a β posterior distribution, that is, $E(p) = (n + 1)/(N + 2)$ and $\text{s.e.m.}(p) = \sqrt{\frac{E(p)(1 - E(p))}{N + 3}}$, in which p is the fraction of bipotent cells, n is the observed number of bipotent cells, and N is the total number of cells assayed.

Growth factor perturbations of erythroid colony formation. CFU-e and BFU-e colony-formation assays in MethoCult (STEMCELL Technologies M3234) were carried out on either freshly isolated adult bone marrow or on fetal liver cells extracted at embryonic day 13.5 from BALB/cJ mice. The following growth factors were tested: MSP/MST1 (R&D Systems 6244-MS-025), recombinant human/mouse WNT5A (R&D Systems 645-WN-010) and recombinant murine IL-17A (PeproTech 210-17). In each experiment, a range of EPO concentrations was tested, with or without additional growth factors (MSP, WNT5A or IL-17A) as indicated in Fig. 5 and Extended Data Fig. 9. In the BFU-e assays, IL-3 (10 ng ml^{-1}) and SCF (50 ng ml^{-1}) were added to the MethoCult in addition to EPO. Each condition was tested in quadruplicate in at least two separate experiments. Colonies were scored on day 3 (for CFU-e), day 4 (for late BFU-e) and day 7 (for early BFU-e) following staining with diaminobenzidine, to highlight haemoglobin expression.

***Il17ra$^{-/-}$* mice.** To generate the *Il17ra$^{-/-}$* line, *Il17ra$^{flox/+}$* mice[68] were bred with CMV-Cre mice (Jackson Laboratory 003465). The generation of the *Il17ra$^-$* allele in the F$_1$ generation of *Il17ra$^{flox/+}$* and CMV-Cre mating pairs was screened by PCR of tail DNA. To remove the *CMV-cre* allele present in the F$_1$ generation, *Il17ra$^{-/+}$CMV-cre$^{+/-}$* mice were outcrossed with B6 mice.

Colony-formation assays with human bone marrow. Human bone marrow mononuclear cells (85,000 cells, STEMCELL Technologies 70001.1) were mixed with 1 ml MethoCult (STEMCELL Technologies H4230) supplemented with EPO (0.05 U ml^{-1}), in the presence or absence of IL-17A (R&D Systems 7955-IL-025). CFU-e colonies were scored from triplicate plates on day 7.

Cell cycle studies. *Flow cytometric cell cycle analysis of bone marrow cells* in vivo. Flow cytometric analyses were carried out as described[47]. In brief, BrdU (100 μl of 10 mg ml^{-1} stock in PBS) was injected intraperitoneally into adult mice 30 min before euthanasia. After collection of bone marrow, cells were immediately placed in cold staining buffer, labelled using a LIVE/DEAD kit (Invitrogen) to identify dead cells and were then fixed and permeabilized. Cell-surface staining for each of the five subsets P1–P5 was carried out as described earlier. Simultaneously, incorporated BrdU was detected using a biotin-conjugated anti-BrdU antibody (Abcam) following mild digestion with DNaseI. DNA content was assayed by labelling with the fluorescent indicator 7-AAD (BD Biosciences). Cells were then analysed for cell-surface labelling, BrdU incorporation and DNA content by flow cytometry.

Cell cycle arrest studies during erythroid differentiation in vitro. Bone marrow cells were collected and immediately enriched for Kit$^+$Lin$^-$TER119$^-$CD71$^-$ cells using magnetic beads, as described earlier. The enriched cell fraction was initially placed in culture in IMDM with 20% FCS and EPO (2 U ml^{-1}), in the presence or absence of aphidicolin (6 μM, Sigma A0781). After 10 h, all the cells were washed three times in culture medium to remove aphidicolin, and returned to culture, which continued for up to a total of 36 h.

At the indicated time points, cell aliquots were taken for RNA extraction followed by qRT–PCR of *Hbb-b1* and *Actb* and for a simultaneous flow cytometric analysis of CD71, TER119 expression and cell cycle status. For the latter, cells were pulsed with BrdU (33 μM) *in vitro* for 25 min before collection, then processed as described earlier for BrdU incorporation, DNA content and cell-surface CD71 and TER119 expression.

Western blot analysis. Bone marrow cells were sorted as described earlier, except that the P1 population was further subdivided into CD71$^{\text{medium}}$ and CD71$^{\text{high}}$ subsets. For negative controls, we used 3T3-L1 cells. For positive controls, 3T3-L1 cells were transduced with the MICD4-GATA1 retrovirus as described[47]. Cell pellets were snap-frozen in liquid nitrogen after sorting.

Cell lysates were quantified using the BCA Protein Assay Kit (Pierce) and separated by SDS–PAGE. PVDF membranes were probed with antibodies against GATA1 (N6, Santa Cruz sc-265), β-actin (Abcam ab8227), MCM5 (Bethyl Laboratories A300-195A-M), MCM6 (Bethyl Laboratories A300-194A), MCM2 (Bethyl Laboratories A300-191A), PCNA (PC10) (Santa Cruz sc-56) and IL-17RA (R&D Systems AF448).

Western blot membranes were quantified using the BIORAD Imaging system and Image Laboratory software.

Intracellular signalling by STAT3 and STAT5. Freshly collected bone marrow cells were enriched for Lin$^-$TER119$^-$ cells using magnetic beads, as described above. The enriched cells were incubated in cytokine-free, low-serum medium (IMDM with 2% FCS) for 3 h. EPO (0.5 U ml^{-1}), IL-17A (20 ng ml^{-1}) or both together was then added to the medium for either 30 or 60 min. Cells were collected, washed with PhosphoWash Buffer[69], stained with a LIVE/DEAD kit (Invitrogen), fixed and permeabilized with Cytofix/Cytoperm Buffer (BD Biosciences 554722) supplemented with 1 mM sodium orthovanadate (Sigma 450243-10G), 1 mM β-glycerophosphate (Sigma G9422-10G) and 1 μg ml^{-1} Microcystin (EMD Millipore 475815-500UG), and Perm/Wash Buffer I

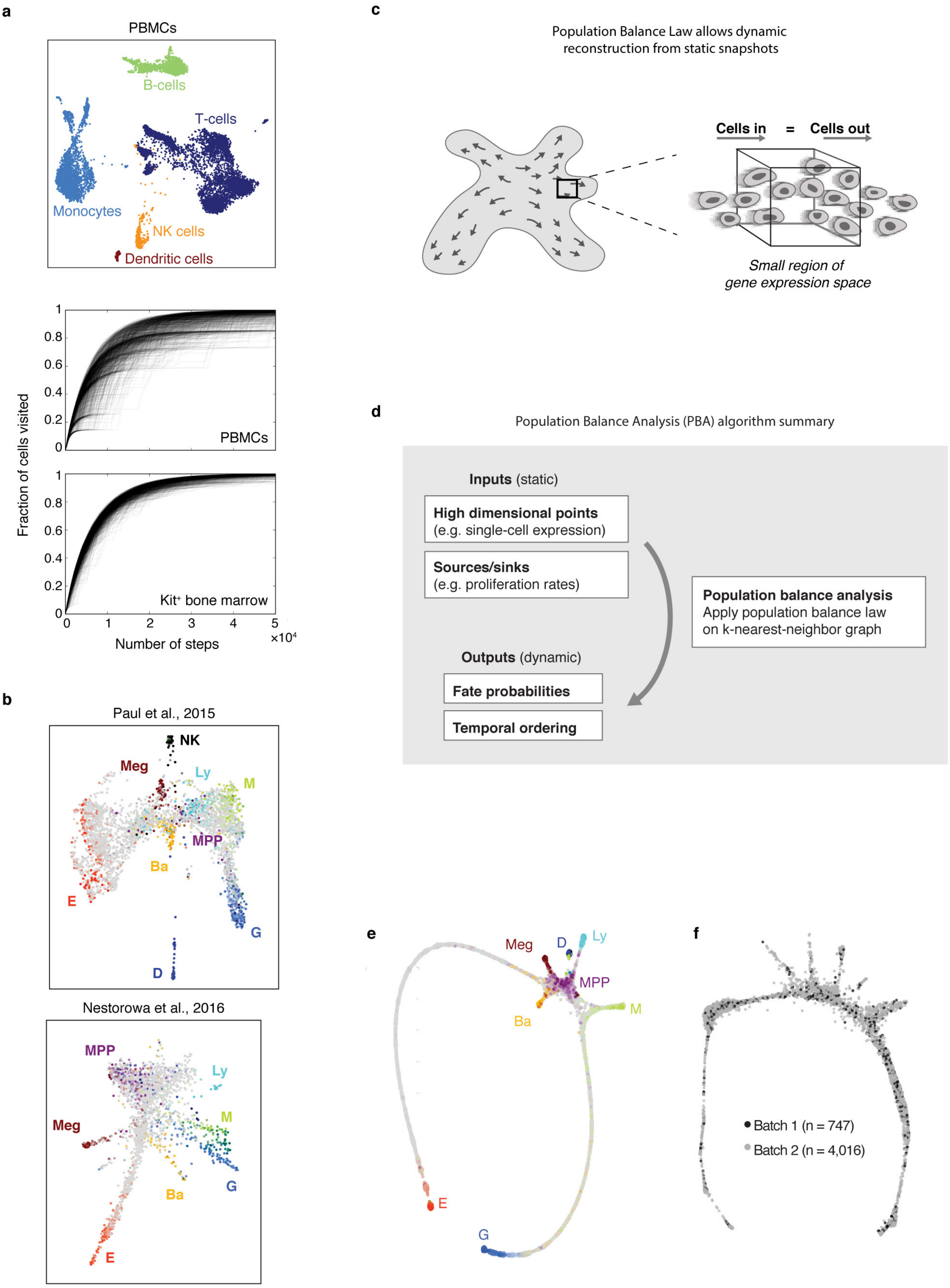
(BD Biosciences 557885), and frozen in freezing medium (90% FCS, 10% DMSO, 1 mM sodium orthovanadate, 1 mM β-glycerophosphate and 1 μg ml^{-1} microcystin). When thawed, cells were re-fixed and permeabilized, incubated with 5% milk and 200 μg ml^{-1} rabbit IgG (modified from ref. 69), and stained with p-STAT3-AF488 (B-7) (Santa Cruz sc-8059 AF488), p-STAT5-AF647 (pY694) (BD Bioscience 612599), CD71–PE/Cy7 (Biolegend 113812), CD55–PE (Biolegend 131804), CD105–Pacific Blue (Biolegend 120412), CD150–BV650 (Biolegend 115931), CD49f–PE/Dazzle 594 (Biolegend 313626), CD41–BV605 (Biolegend 133921), CD117 (Kit)–APC/H7 (BD Bioscence 560185), strepavidin–AF700 (Invitrogen S21383) and DAPI. Analysis was performed on an LSRII FACS analyser.

Code availability. Python scripts are described in the PBA section, and Supplementary Data 1 contains the input data files and code for running PBA on the bone marrow and fetal liver datasets. Code is available at https://github.com/indrops/indrops, https://github.com/AllonKleinLab/SPRING and https://github.com/AllonKleinLab/PBA.

Data availability. Sequence data that supports the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE89754. An interactive tool for the interpretation of these data is available at https://kleintools.hms.harvard.edu/paper_websites/tusi_et_al/.

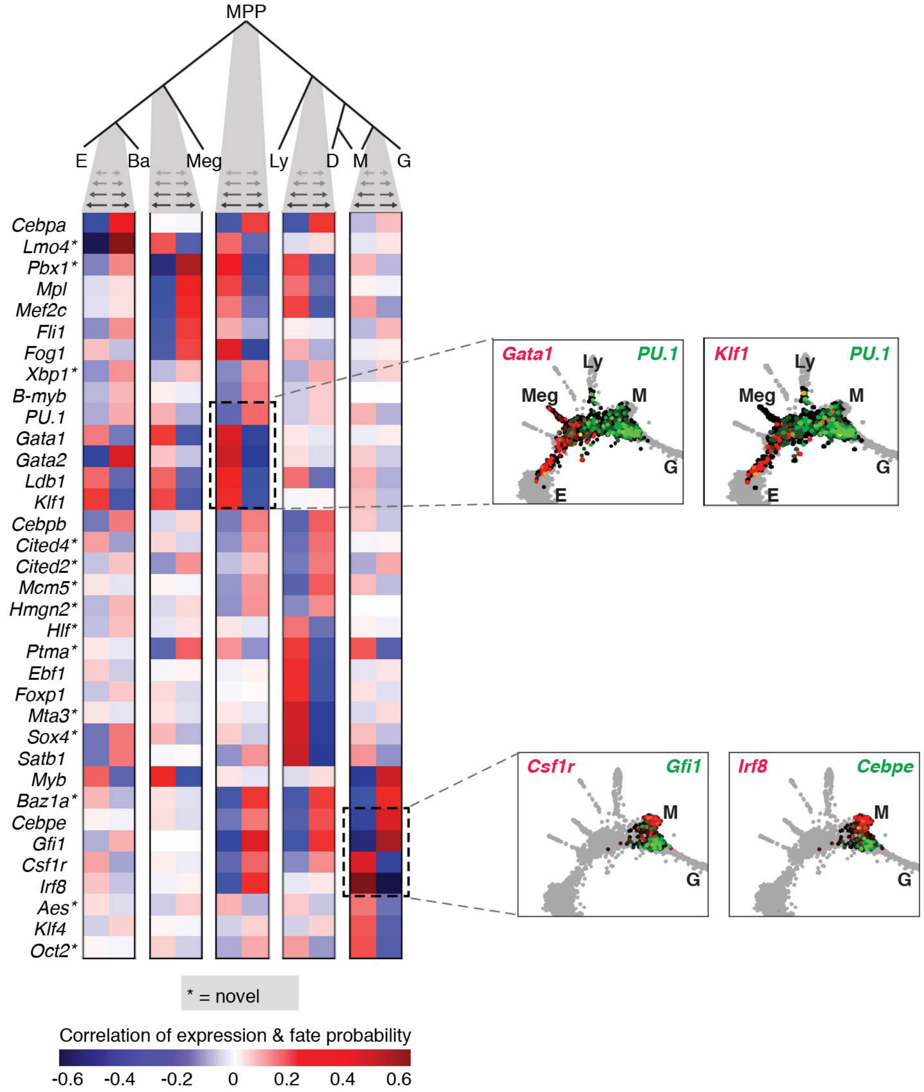
Source Data files are provided for Figs 2c–e, 3b, 5b–d, 6f, Extended Data Figs 3a, 4c, 5a, b, 7b, 9b, 10e, f–h and for all immunoblots (Supplementary Fig. 1).

53. Zilionis, R. *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protocols* **12,** 44–73 (2017).
54. Ester, M., Kriegel, H., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd International Conference on Knowledge Discovery and Data Mining* (Eds Simoudis, E. *et al.*) 226–231 (AAAI, 1996).
55. Daszykowski, M., Walczak, B. & Massart, D. L. Looking for natural patterns in data: Part 1. Density-based approach. *Chemomtr. Intell. Lab. Syst.* **56,** 83–92 (2001).
56. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).
57. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161,** 1202–1214 (2015).
58. Weinreb, C., Wolock, S. & Klein, A. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btx792 (2017).
59. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18,** 507–522 (2011).
60. Heng, T. S. *et al.* The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9,** 1091–1094 (2008).
61. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
62. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26,** 2438–2444 (2010).
63. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85,** 54–61 (2015).
64. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43,** D1140–D1144 (2015).
65. Shekhar, K. *et al.* Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166,** 1308–1323.e1330 (2016).
66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289–300 (1995).
67. Tusi, B. K. & Socolovsky, M. Novel FACS strategy for identification of early hematopoietic progenitors including BFU-e, CFU-e and erythroid-biased MPPs *Protoc. Exch*. http://doi.org/10.1038/protex.2018.031 (2018).
68. El Malki, K. *et al.* An alternative pathway of imiquimod-induced psoriasis-like skin inflammation in the absence of interleukin-17 receptor a signaling. *J. Invest. Dermatol.* **133,** 441–451 (2013).
69. Porpiglia, E., Hidalgo, D., Koulnis, M., Tzafriri, A. R. & Socolovsky, M. Stat5 signaling specifies basal versus stress erythropoietic responses through distinct binary and graded dynamic modalities. *PLoS Biol.* **10,** e1001383 (2012).

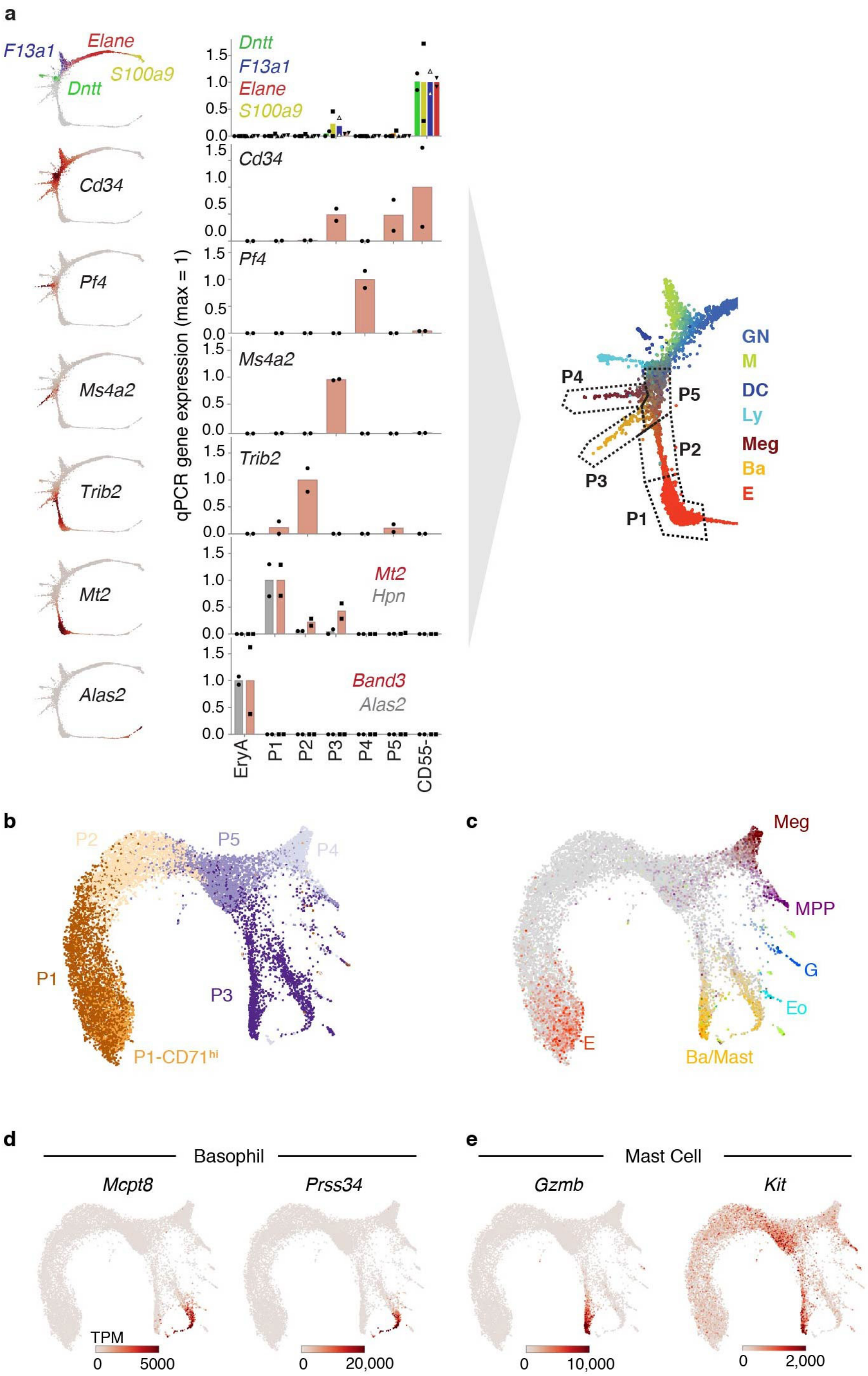


Extended Data Figure 1 | See next page for caption.

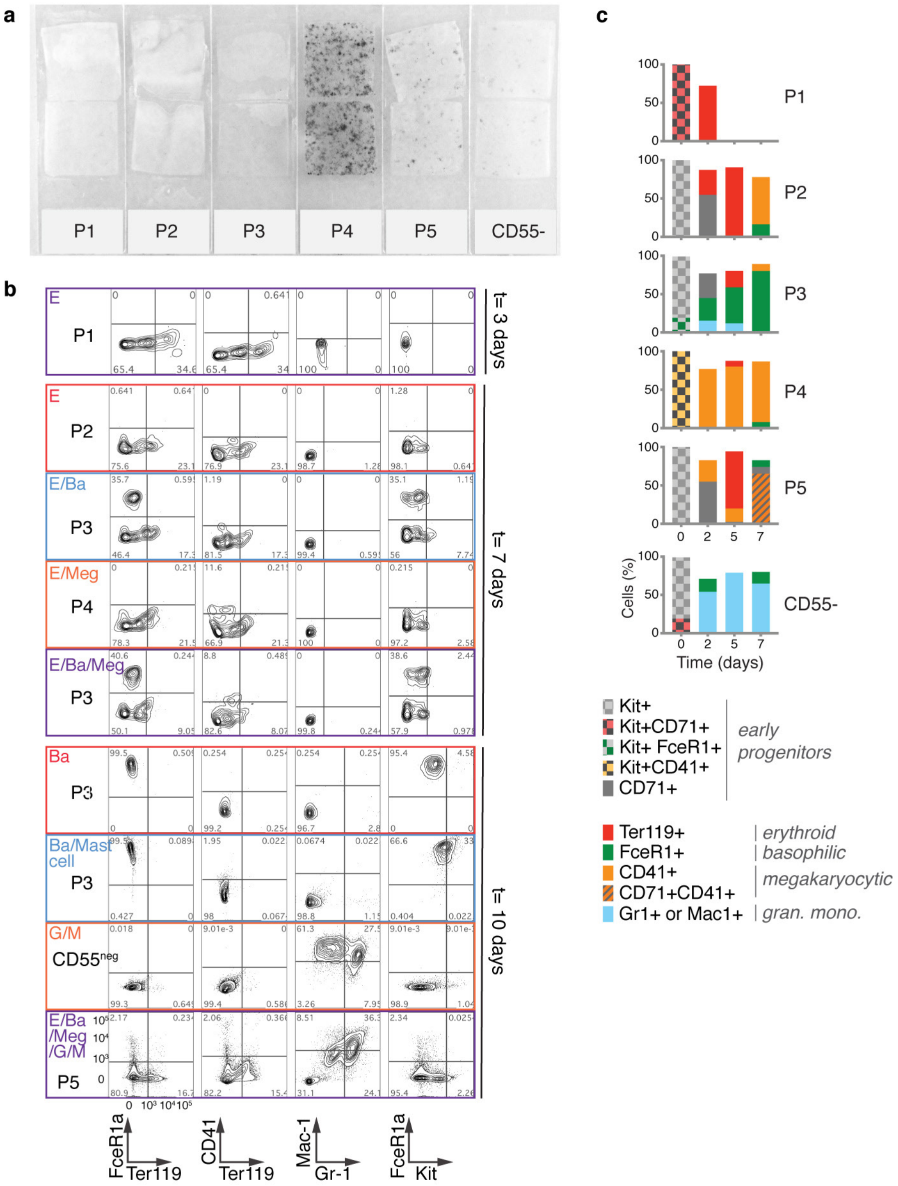
Extended Data Figure 1 | scRNA-seq of Kit$^+$ haematopoietic progenitors for prediction of the early haematopoietic hierarchy. **a**, Top, SPRING plot of 7,959 human peripheral blood mononuclear cells (PBMCs) from 10X Genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc8k). Clusters were generated by performing spectral clustering on the underlying kNN graph and annotated on the basis of marker genes. NK, natural killer. Random walks over kNN graphs for the PBMC (middle) and Kit$^+$ bone marrow (bottom) datasets. Each plot shows the fraction of nodes (cells) visited for 1,000 simulated random walks. **b**, Top, SPRING plot of 2,855 Lin$^-$Kit$^+$SCA1$^-$ mouse HPCs from a previously published dataset[12]. Bottom, SPRING plot of 1,656 cells from three mouse haematopoietic progenitor populations (Lin$^-$Kit$^+$Sca1$^-$, Lin$^-$Kit$^+$SCA1$^+$, and Lin$^-$Kit$^+$SCA1$^+$FLK2$^-$CD34$^+$) from a previously published dataset[15]. Coloured (non-grey) cells indicate expression of lineage-specific genes (see Supplementary Table 7). **c**, Schematic of the population balance law, which relates the dynamic velocities of cells to the distribution of states they are in at a moment in time. The law states that in steady state, after accounting for cell division and loss, the flux of cells entering any region of gene expression space equals the flux out of that region. **d**, Flow diagram of the inputs and outputs of the PBA algorithm. The population balance law is applied to inputs that include single-cell expression data and estimates of cell proliferation and loss rates at each point in gene expression space; inferred outputs include cell dynamics such as fate probabilities and pseudo-temporal ordering. **e**, SPRING plot of bone marrow Kit$^+$ cells (Fig. 1) constructed using only the PBA-predicted fate probabilities and differentiation ordering as inputs ($n = 4,763$ cells from one inDrops experiment). Coloured cells indicate expression of lineage-specific genes as in Fig. 1b. **f**, SPRING plot of bone marrow Kit$^+$ cells (Fig. 1), with cells coloured by library preparation batch.



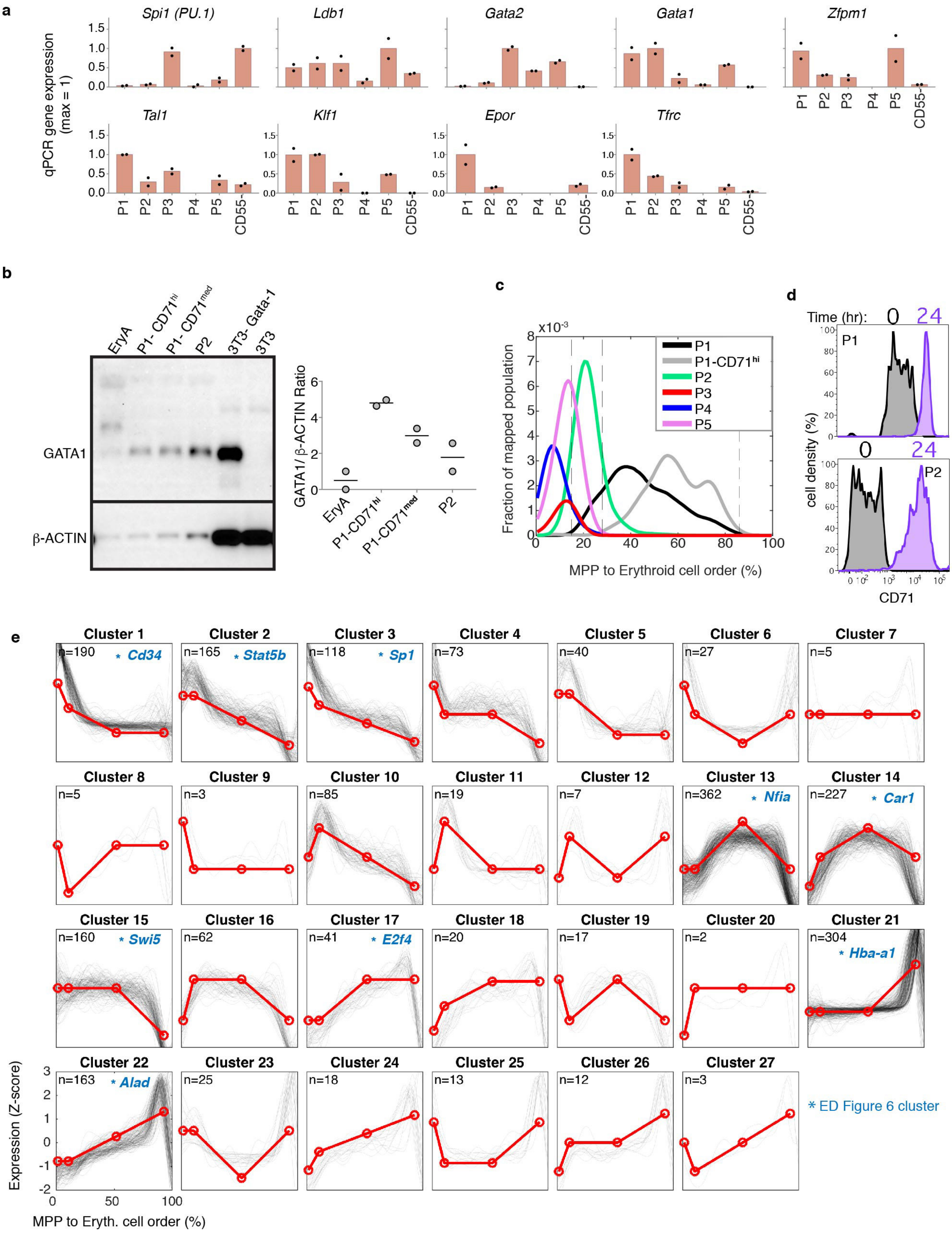
Extended Data Figure 2 | Predicting key regulators at haematopoietic choice points. Candidate regulators of fate choice, identified by ranking transcription factors and transmembrane receptors by their correlation with PBA-predicted fate probabilities at key choice points in haematopoiesis. Top-ranked genes are shown; these include many canonical regulators. Candidate genes that have not been previously reported are marked with asterisks. Several candidates participate in more than one fate choice. Insets show SPRING plots coloured by expression of representative genes. *Fog1* is also known as *Zfpm1*, *B-myb* is also known as *Mybl2*, *PU.1* is also known as *Spi1* and *Oct2* is also known as *Slc22a2*.



Extended Data Figure 3 | Mapping HPC subsets P1–P5 to the Kit$^+$ SPRING plot using qRT–PCR and scRNA-seq. a, Subpopulations P1–P5 map onto specific regions of the SPRING plot. SPRING plot heat maps for a panel of marker genes (left) and corresponding measured expression for each of the marker genes by qRT–PCR (middle), performed on sorted cell subsets P1–P5, and on EryA (cells undergoing ETD[2]). Probable mapping of each of the P1–P5 subpopulations on the basis of qRT–PCR is shown on the SPRING plot (right). Bars represent the mean of two independent experiments (circles, triangles or squares). Expression is shown normalized to *Actb* mRNA. **b**, **c**, SPRING plot of single-cell transcriptomes from freshly sorted P1–P5 subsets (Fig. 2a, b). Cells are coloured on the basis of sorted subpopulation (**b**) or the expression of lineage-specific marker genes (**c**) (Supplementary Table 7). **d**, **e**, SPRING plots of P1–P5 subpopulation cells, coloured on the basis of expression of basophil (**d**) and mast cell (**e**) marker genes. The larger number of cells in the P3 region of the graph resolves a split between the two lineages that was not observable in the original Kit$^+$ dataset.

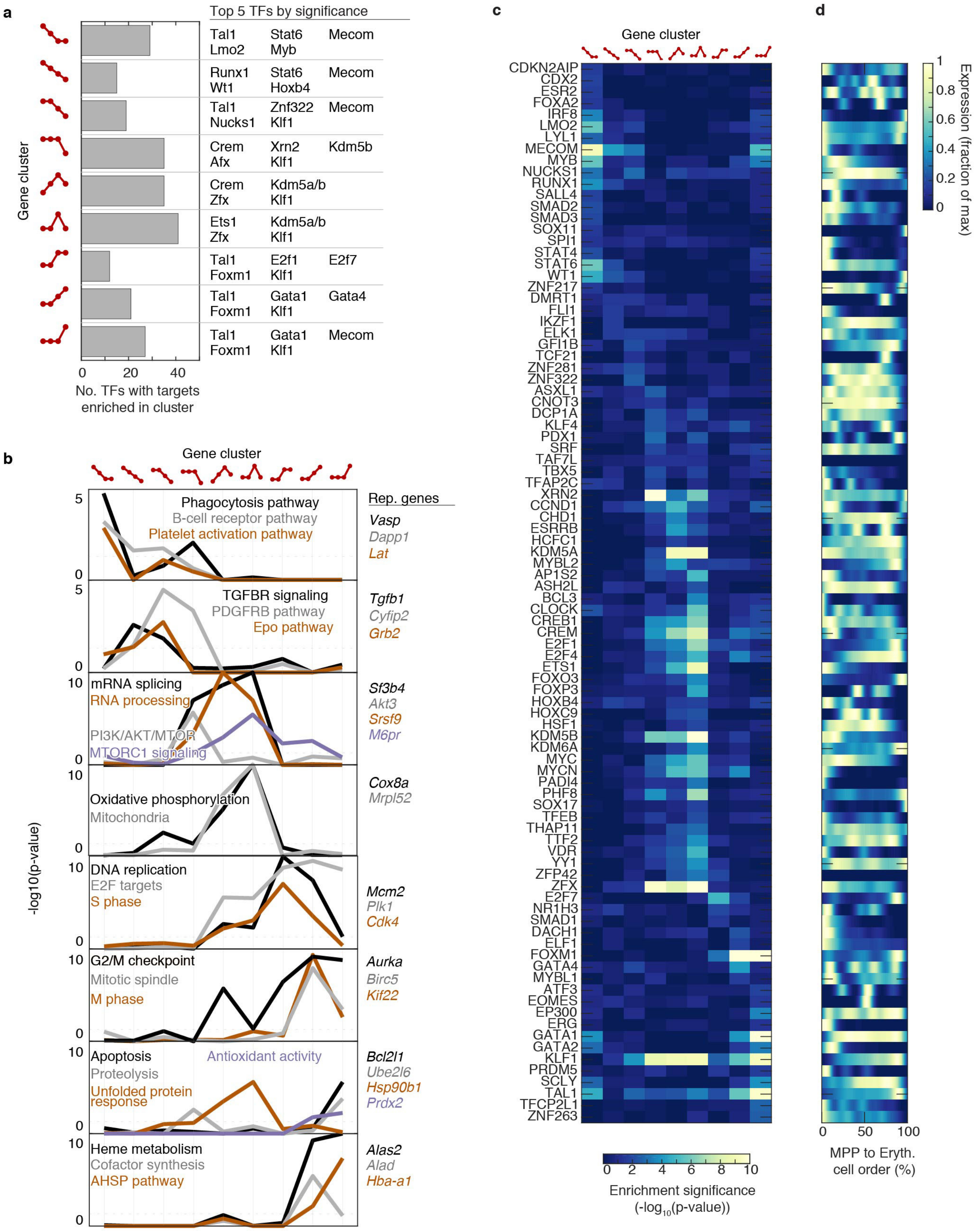


Extended Data Figure 4 | Validation of PBA predictions. a, Megakaryocytic colonies from sorted subsets P1–P5 and from Kit$^+$CD55$^-$ cells, stained for the megakaryocytic marker acetylcholinesterase. Duplicate cultures are shown; representative of two independent experiments. **b**, Representative flow cytometry plots to assay fate output of single cells in liquid culture (see Fig. 3, including Fig. 3a for experimental design). Each row corresponds to a single clone, with the left column indicating the source subset (P1–P5, CD55$^-$) of the clone and the cell type(s) produced, as inferred from the FACS plots in the remaining columns. These data are representative of 1,158 single-cell clonal cultures, pooled from three independent sorting experiments (complete dataset is shown in Fig. 3b). **c**, Bulk liquid cultures of freshly sorted P1–P5 subsets and Kit$^+$CD55$^-$ cells in the presence of EPO and a mixture of cytokines supporting myeloid progenitors. On the indicated days, cells were labelled with antibodies as indicated and analysed by flow cytometry.



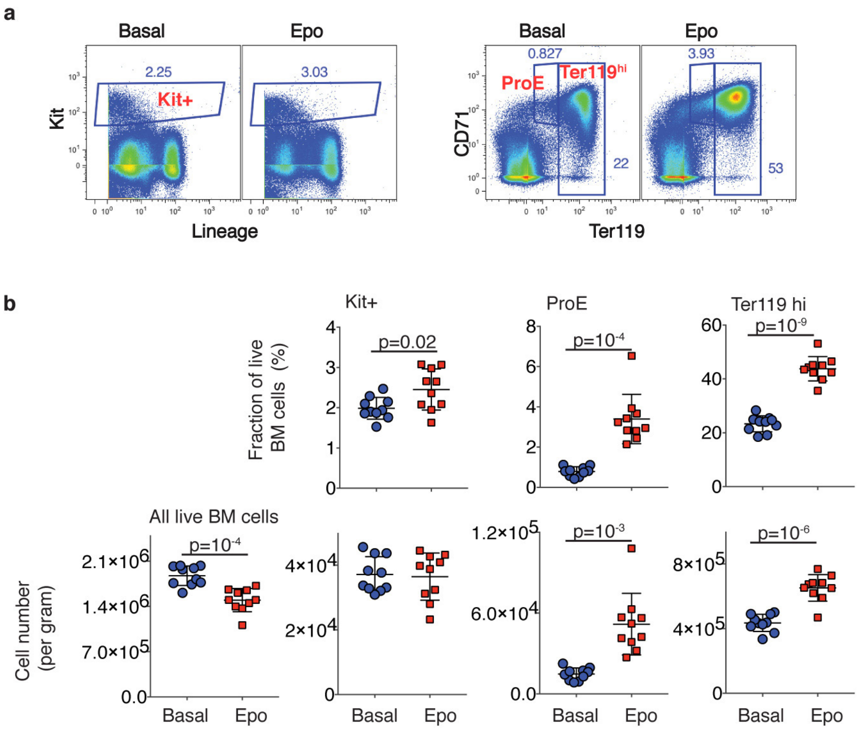
Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | The early erythroid trajectory. a, qRT–PCR for expression of established erythroid regulators in sorted P1–P5 subsets. Expression of each gene is normalized to *Actb*. Bars represent the mean of two independent experiments (circles). **b**, Left, representative western blot of GATA1 expression in sorted P1 (subdivided into CD71$^{\text{medium}}$ and CD71$^{\text{high}}$ subsets), P2 and EryA (CD71^{+}TER119^{+}FSC$^{\text{high}}$ cells, representative of ETD) cells. 3T3-GATA1, positive control 3T3 cells virally transduced with a GATA1 expression vector; untransduced 3T3 cells were used as negative control. Right, quantification of GATA1 expression (mean) by densitometry. Data points are from two independent western blots. For gel source data, see Supplementary Fig. 1. **c**, Density of FACS subsets P1–P5 along the erythroid trajectory. Single-cell transcriptomes from each subset were mapped to their most similar counterparts in the Kit^{+} data (Fig. 2a, b; cell numbers analysed for each subset are indicated in Fig. 2b). Shown here is the fraction of mapped cells after smoothing with a Gaussian kernel. Also included are CD71$^{\text{high}}$ P1 cells, constituting cells with the 30% highest CD71 expression in that subset ($n = 752$ cells post-filter). **d**, Distribution of CD71 expression in P1 (top) and P2 (bottom) cells immediately after sorting (grey) and after 24 h of *in vitro* differentiation (lavender). Data are representative of two independent experiments. **e**, Dynamically varying genes along the MPP-to-erythroid axis were clustered on the basis of their behaviour across three transition points. At each transition, gene expression is increased, decreased or unchanged, giving a total of 27 potential dynamic patterns across all three transitions, shown in red. The number of genes corresponding to each dynamic pattern is noted, and *z*-score-normalized expression traces for each individual gene are shown in black. Selected clusters were further analysed in Extended Data Fig. 6 and are marked with an asterisk and a representative gene.

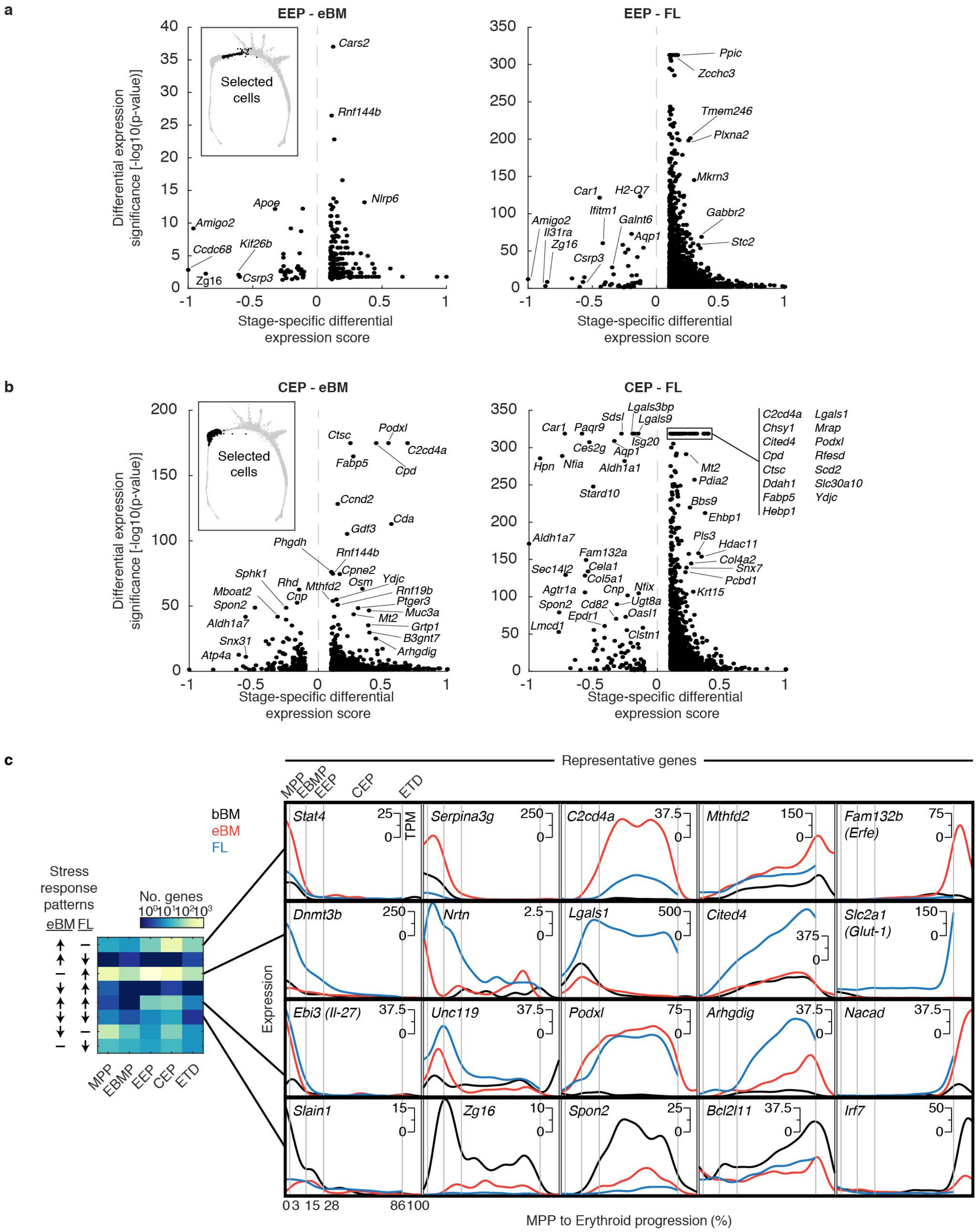


Extended Data Figure 6 | See next page for caption.

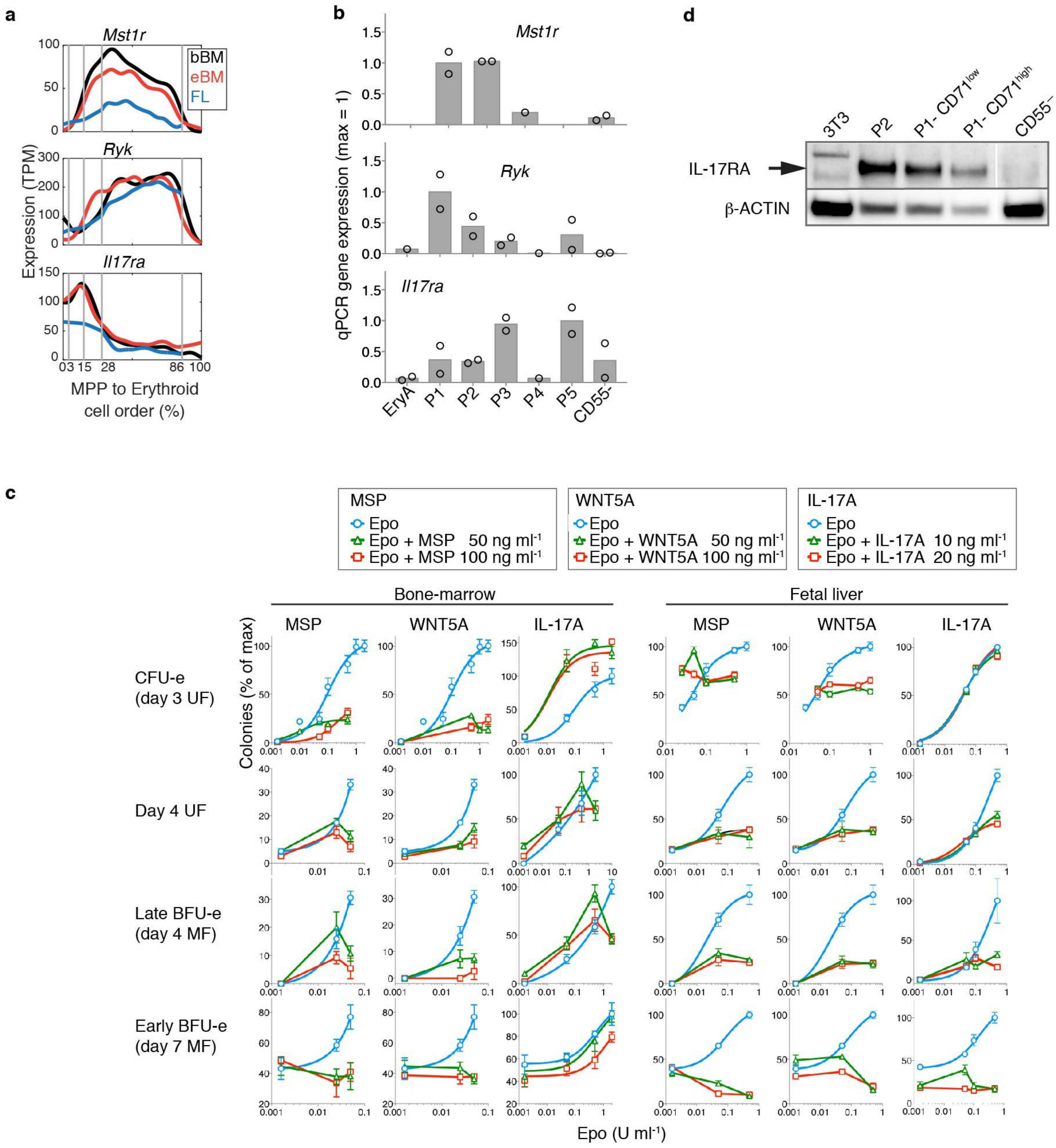
Extended Data Figure 6 | Gene set enrichment on dynamic gene clusters in early erythroid differentiation. a–d, Nine key dynamic gene clusters along the MPP-to-erythroid progression (Extended Data Fig. 5e) are analysed further for Gene Ontology terms. Each cluster is identified by its dynamic pattern with a cartoon of nodes and edges. Each node represents a progenitor stage (in order, MPP, EBMP/EEP, CEP and ETD), connected to the next stage by an edge that either goes up (for increased expression) or down (for decreased expression) (see Extended Data Fig. 5e). **a**, Number and identity of transcription factors (TFs) whose targets are enriched in the dynamic clusters, as predicted by chromatin immunoprecipitation methods (ChIP-X) experiments[62]. **b**, Significance of enrichment for signalling pathways and Gene Ontology gene sets in the dynamic gene clusters (hypergeometric test with Benjamini–Hochberg correction for multiple hypothesis testing). Rep., representative. **c**, **d**, Enrichment of transcription factor targets. **c**, Heat map ($-\log_{10}$ of P value of hypergeometric test with Benjamini–Hochberg correction for multiple hypothesis testing) of target gene enrichment for transcription factors (rows) with targets significantly enriched ($P < 0.05$) in at least one of the nine dynamic clusters (columns, labelled on top) highlighted in Extended Data Fig. 5e. Note that the transcription factor targets shown are based on previous ChIP-X experiments[62] and it is possible that unappreciated transcription factor targets occur in early erythropoiesis. **d**, Gene expression traces over the erythroid trajectory for the transcription factors from **c**. Rows match those in **c**.



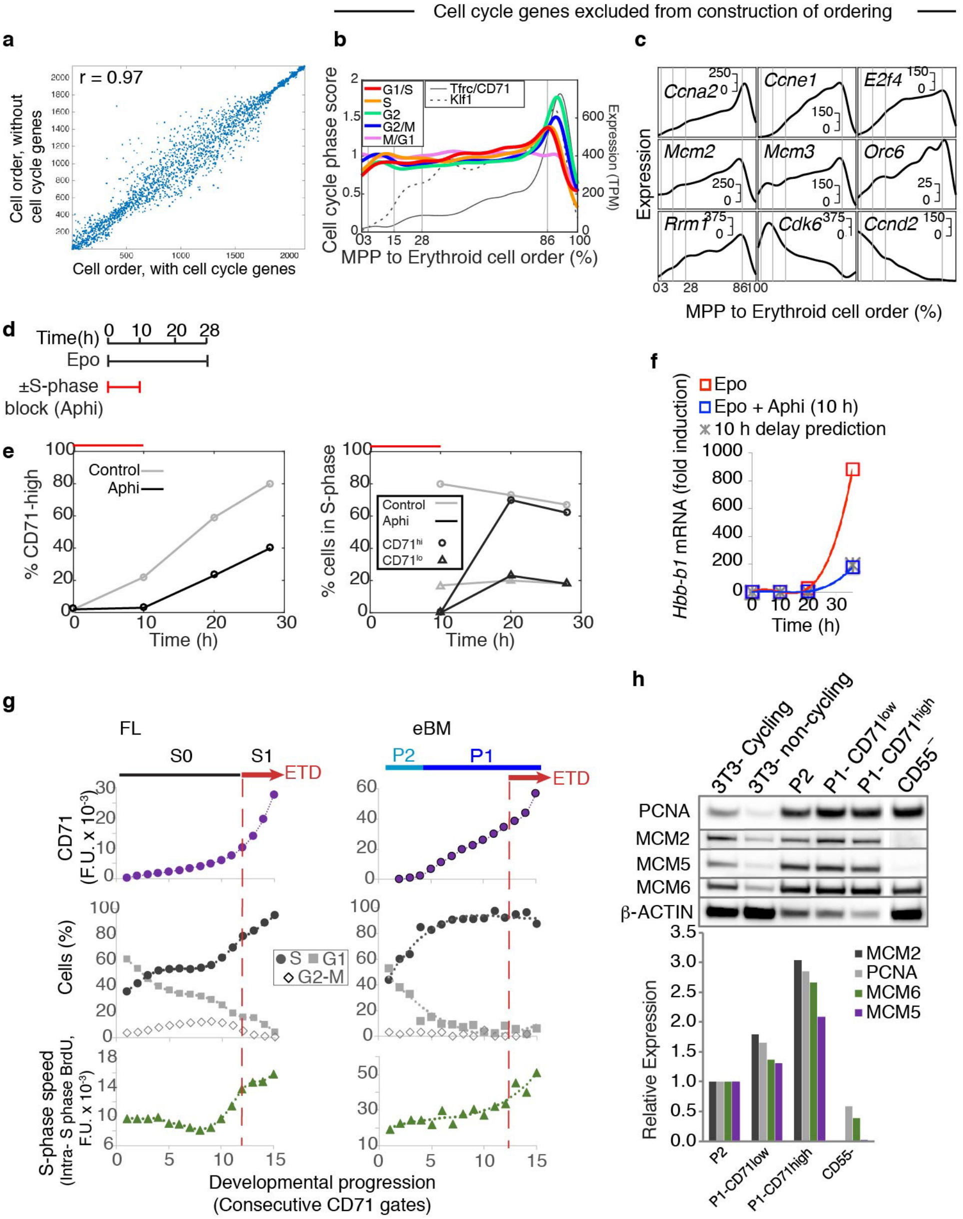
Extended Data Figure 7 | Quantification of absolute Kit$^+$ cell number in bone marrow after *in vivo* administration of EPO. a, **b**, Eight-week-old female mice were injected with either EPO (100 U per 25 g body weight) or saline (basal), once per day for two days. Bone marrow was collected at 48 h. Viable (trypan blue-negative) cells were counted using a TC20 automated cell counter (BIORAD) and stained for Kit, TER119 and CD71 and lineage markers. Data are from two independent experiments, with five mice analysed individually for each group (basal or EPO) in each experiment. **a**, Representative flow cytometric analysis of either basal or EPO-stimulated bone marrow, gating on Kit$^+$Lin$^-$ cells (left) or on proerythroblasts (ProE) and TER119^{high} cells (right; ProE and TER119^{high} cells are sequential stages of ETD). **b**, Data summary (mean ± s.d.) for all mice (ten in each group). Top, the fraction of all bone marrow cells for each of the flow cytometric gates defined in **a**. Bottom, the absolute cell count in adult bone marrow for subsets defined in each flow cytometric gate, or for the total number of bone marrow cells. Significant ($P < 0.05$) P values are shown (two-tailed t-test, unequal variances).



Extended Data Figure 8 | Identification of stage-specific differential gene expression during the erythroid stress response. a, Identification of genes that are differentially expressed in EEP cells of either EPO-stimulated bone marrow (eBM, left) or fetal liver (FL, right) samples, compared with basal bone marrow (bBM). One single-cell inDrop experiment per condition. *P* values were calculated using a binomial test for differential expression (see Methods) and measure the significance of the expression difference. The specific enrichment score (see Methods) measures the degree to which the differential expression is specific to this region of interest (EEPs); positive scores correspond to region-specific upregulation, and negative to region-specific downregulation. Selected genes are highlighted. **b**, The analysis in **a** applied to the CEP stage. One single-cell inDrop experiment per condition. **c**, Stage-specific differential gene expression during stress, comparing EPO-stimulated and fetal liver samples. The heat map shows the number of differentially expressed genes at each stage that show similar or different patterns of upregulation and downregulation in fetal liver and EPO-stimulated bone marrow samples. Representative gene traces are shown on the right.



Extended Data Figure 9 | Localized gene expression and functional response of the erythroid lineage to stimulation of MST1, RYK and IL-17RA. **a**, **b**, Predicted expression pattern (**a**) and confirmation by qRT–PCR (**b**) for *Mst1r*, *Ryk* and *Il17ra* in basal bone marrow. In **a**, traces show the smoothed scRNA-seq gene expression of cells from the basal bone marrow (bBM), fetal liver (FL) and EPO-stimulated bone marrow (eBM) samples, arranged along the erythroid trajectory. qRT–PCR data represent the mean (bars) of two independent experiments (circles). **c**, Complete results for CFU-e and BFU-e colony-formation assays in methylcellulose, supporting the data shown in Fig. 5. Curves show colony numbers in the presence of increasing concentrations of either EPO, or EPO with a ligand (MSP, WNT5A or IL-17A). Error bars show s.d. of two independent experiments, with four replicates per experiment. Where appropriate, data were fitted to a dose–response curve with a Hill coefficient of one. **d**, Western blot analysis shows that IL-17RA expression peaks in EEP P2 cells, dropping in CEP P1 cells and in the granulocytic branch (which contributes most of the CD55⁻ cells), consistent with the SPRING plots in Fig. 5a. The western blot is representative of two independent experiments. For gel source images, see Supplementary Fig. 1.



Extended Data Figure 10 | Cell ordering independently of cell cycle genes, and evidence of an S-phase-dependent CEP-to-ETD transition in bone marrow erythropoiesis. a, The computational ordering of cells from MPP to ETD is not sensitive to the inclusion or exclusion of annotated cell cycle genes (cell ordering correlation is $R = 0.97$). **b**, **c**, Reproduction of Fig. 6b, c after the exclusion of cell cycle genes shows that the computationally inferred expression dynamics of cell cycle genes during EEP to CEP differentiation are not sensitive to the inclusion or exclusion of annotated cell cycle genes when ordering cells. **d–f**, Activation of ETD is dependent on the S phase. **d**, Schematic of experiments shown in **e** and **f** that test the link between S-phase progression and the CEP-to-ETD transition. $\text{Kit}^{+}\text{Lin}^{-}\text{CD71}^{-}$ bone marrow cells were cultured in the presence of EPO for 28 h, with or without the presence of the DNA polymerase inhibitor aphidicolin (Aphi) for the first 10 h. **e**, $\text{Kit}^{+}\text{Lin}^{-}\text{CD71}^{-}$ bone marrow cells require the S phase to upregulate CD71, an early event in ETD. Left, $\text{CD71}^{\text{high}}$ cells fail to appear in the first 10 h if cells are exposed to aphidicolin; they appear as soon as aphidicolin is removed from the medium. Right, cell cycle analysis of the same cells shows that aphidicolin prevented S-phase progression; aphidicolin removal was followed by a full recovery of S-phase progression, with a high fraction of $\text{CD71}^{\text{high}}$ cells in S phase. Data are representative of three independent experiments. **f**, Aphidicolin exposure for 10 h delays induction of β-globin (*Hbb-b1*) by 10 h. Data are representative of two independent experiments. **g**, CD71 expression (top), cell cycle phase distribution (middle), and intra-S-phase DNA synthesis rate (bottom), for consecutive FACS gates of increasing CD71 in the early stages of erythropoiesis from the fetal liver (left, representative of four independent experiments) and EPO-simulated bone marrow (right, representative of two independent experiments) samples. See Fig. 6e, f for similar analysis of basal bone marrow samples. **h,** Western blots (top) and quantification by densitometry (bottom) showing an increase in S-phase proteins during progression from EEP (P2) to early CEP ($\text{P1-CD71}^{\text{low}}$) and late CEP ($\text{P1-CD71}^{\text{high}}$). Control 3T3 cells were either cycling or contact-inhibited (non-cycling), as indicated. Western blots are representative of three independent experiments. For gel source images, see Supplementary Fig. 1.

doi:10.1038/nature25762

Extreme disorder in an ultrahigh-affinity protein complex

Alessandro Borgia[1]*, Madeleine B. Borgia[1]*, Katrine Bugge[2]*, Vera M. Kissling[1], Pétur O. Heidarsson[1], Catarina B. Fernandes[2], Andrea Sottini[1], Andrea Soranno[1,3], Karin J. Buholzer[1], Daniel Nettels[1], Birthe B. Kragelund[2], Robert B. Best[4] & Benjamin Schuler[1,5]

Molecular communication in biology is mediated by protein interactions. According to the current paradigm, the specificity and affinity required for these interactions are encoded in the precise complementarity of binding interfaces. Even proteins that are disordered under physiological conditions or that contain large unstructured regions commonly interact with well-structured binding sites on other biomolecules. Here we demonstrate the existence of an unexpected interaction mechanism: the two intrinsically disordered human proteins histone H1 and its nuclear chaperone prothymosin-α associate in a complex with picomolar affinity, but fully retain their structural disorder, long-range flexibility and highly dynamic character. On the basis of closely integrated experiments and molecular simulations, we show that the interaction can be explained by the large opposite net charge of the two proteins, without requiring defined binding sites or interactions between specific individual residues. Proteome-wide sequence analysis suggests that this interaction mechanism may be abundant in eukaryotes.

In the traditional paradigm of structural biology, intermolecular interactions are thought to be encoded in complementary shapes and noncovalent forces between folded biomolecules. However, it has become increasingly clear that many proteins involved in cellular interactions are fully or partially unstructured under physiological conditions[1,2]. Some of these intrinsically disordered proteins (IDPs) form well-defined 3D structures on binding to their targets[1]; in other complexes, parts of the IDP remain disordered. A broad spectrum of these protein complexes with differing degrees of disorder is known[3]. In some cases, a well-defined and structured binding interface is formed in the bound state, and only some loops or the chain termini remain disordered. In other cases, one of the binding partners remains almost completely unstructured in the complex, and its multiple binding motifs dynamically interact with the folded partner. Examples of the latter complexes include interdomain interactions in the cystic fibrosis transmembrane regulator[4]; the cyclin-dependent kinase inhibitor Sic1 in yeast binding to the substrate recognition subunit of its ubiquitin ligase subunit Cdc4[5]; the tail of the human sodium/proton exchanger 1 with the extracellular signal-regulated kinase ERK2[6]; and nuclear transport receptors interacting with nucleoporins[7]. The underlying multivalent binding enables unique regulatory mechanisms[8] and can mediate the formation of liquid–liquid phase separation[9], indicating the emergence of new modes of biomolecular interactions.

We have discovered a pair of proteins that constitute an extreme case of a highly unstructured protein complex with physiological function. One of the binding partners is the linker histone H1.0 (H1), which is involved in chromatin condensation by binding to nucleosomes[10,11]; this protein is largely unstructured[12] and highly positively charged, with two disordered regions that flank a small folded globular domain (Fig. 1, Extended Data Table 1). The other partner is the abundant nuclear protein prothymosin-α (ProTα), which is a fully unstructured, highly negatively charged IDP[13,14] involved in chromatin remodelling[15], transcription, cellular proliferation and apoptosis[16]. ProTα acts as a linker histone chaperone by interacting with H1 and increasing its mobility in the nucleus[17]. Here we show that ProTα and H1 bind to one another with very high affinity, but that both proteins fully retain their structural disorder. By integrating experimental techniques and molecular simulations, we obtain a detailed model of this highly disordered and dynamic protein complex, which presents a previously undescribed paradigm of biomolecular binding.

A highly unstructured protein complex

The binding of H1 to ProTα has been demonstrated both *in vitro*[18] and *in vivo*[17]. However, the high net charge, low hydrophobicity and pronounced disorder in the free proteins raise the question of how much structure is formed when they interact. We used circular dichroism and nuclear magnetic resonance (NMR) spectroscopy to investigate the formation of secondary and tertiary structure in ProTα and H1, separately and in complex with one another. The circular dichroism spectra of unbound ProTα and H1 reflect the low secondary structure content of each individual IDP, except for the small helix-turn-helix domain of H1[13,19,20] (Fig. 1c). Notably, the circular dichroism spectrum of an equimolar mixture of the two proteins can be explained by the simple sum of the individual spectra, indicating that complex formation entails minimal changes in average secondary structure content.

To obtain residue-specific information, we employed NMR spectroscopy. ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectra of the individual proteins exhibit low dispersion of the ^{1}H chemical shifts, as expected for IDPs[14,21–23] (Fig. 1e, f). Only the globular domain of H1, which is stably folded even in isolation (Extended Data Fig. 1), shows the large dispersion of resonances characteristic of tertiary structure[23,24] (Fig. 1g). Remarkably, the overall peak dispersion remains unchanged on complex formation, confirming that no pronounced tertiary structure is formed on binding. Nevertheless, small but clearly

[1]Department of Biochemistry, University of Zurich, 8057 Zurich, Switzerland. [2]Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science and Integrative Structural Biology at University of Copenhagen (ISBUC), Department of Biology, University of Copenhagen, 2200 Copenhagen N, Denmark. [3]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [4]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA. [5]Department of Physics, University of Zurich, 8057 Zurich, Switzerland.

*These authors contributed equally to this work.

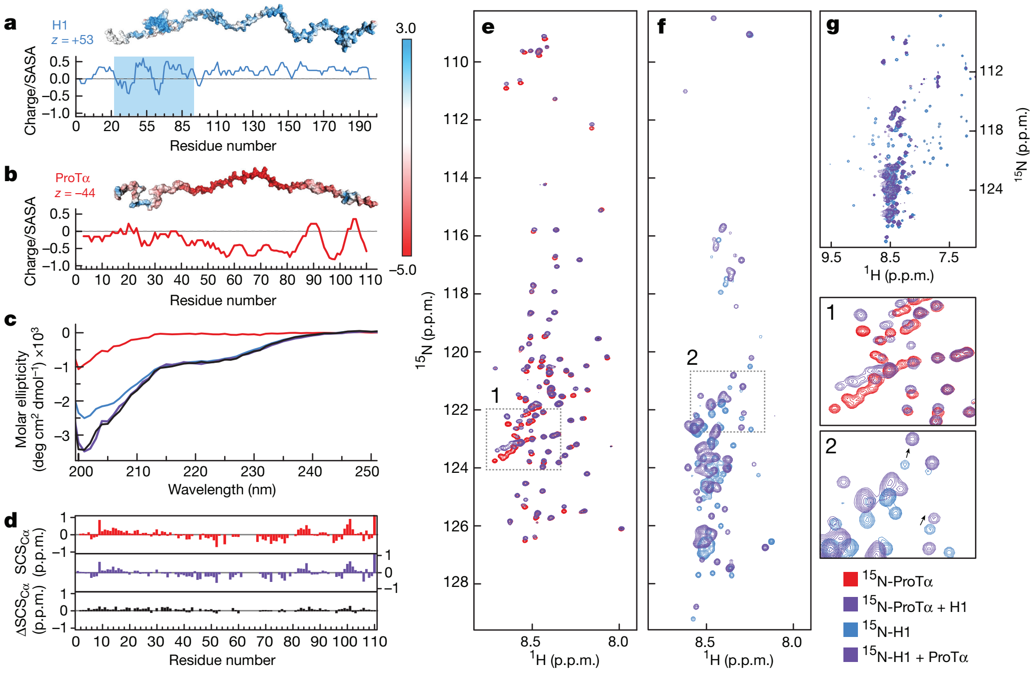


Figure 1 | ProTα and H1 remain unstructured upon binding. **a**, **b**, Extended configurations of H1 (**a**) and ProTα (**b**), net charges (*z*), and surface electrostatic potentials with colour scale (units in k_BT per e$^-$). For the globular domain of H1, only residues with a solvent-accessible surface area (SASA) larger than 0.5 nm^2 are included, and are indicated by blue shading (see Extended Data Table 1). **c**, Far-UV circular dichroism spectra of ProTα (red), H1 (blue), the H1–ProTα mixture (purple) and their calculated sum (black) at 5 μM for each protein; curves are the mean of $n = 60$ individual spectra; $n = 2$ repeats of this measurement yielded consistent results. **d**, Cα secondary chemical shifts (SCS$_{C\alpha}$) of ProTα free (red), in complex with H1 (purple) and their differences (ΔSCS$_{C\alpha}$, black). **e**, ^{1}H–^{15}N HSQC spectra of ^{15}N-ProTα in the absence (red) and presence (purple) of unlabelled H1; $n = 5$ repeats of this measurement yielded consistent results. **f**, ^{1}H–^{15}N HSQC spectra of ^{15}N-H1 in the absence (blue) and presence (purple) of unlabelled ProTα ($n = 2$), with zoomed-in regions corresponding to boxes in **e** (1) and **f** (2); arrows in 2 indicate the shift for selected resonances (see also Extended Data Fig. 3b). **g**, H1 spectra from **f** shown at a lower contour level.

detectable peak shifts observed for ProTα and H1 indicate changes in the average chemical environment of the corresponding residues, as expected on interaction with the large opposite charge of the other IDP. For ProTα, 95% of the amide backbone nuclei could be assigned (Extended Data Fig. 2), which enabled a residue-specific analysis: the Cα secondary chemical shifts[25] of ProTα show no evidence for the induction of persistent or transiently populated secondary structure on complex formation (Fig. 1d). The pronounced overlap in the NMR spectra of the unstructured parts of H1 precluded residue-specific assignments, but the clusters of Hα–Cα peaks in the ^{1}H–^{13}C HSQC spectrum from the lysine-rich disordered regions of H1 do not exhibit detectable chemical shift perturbations on titration with ProTα, and additional resonances do not emerge (Extended Data Fig. 3e, f). We therefore have no indications of changes in secondary structure content in H1 on ProTα binding.

The lower intensity of the resonances corresponding to the H1 globular domain (Fig. 1f, g, Extended Data Fig. 3) is likely to originate from the faster transverse (T_2) relaxation of structured, compared to unstructured, regions; additionally, tumbling of the globular domain is decelerated by the drag of the unstructured regions in which it is embedded[26]. On complex formation, the intensity of many H1 and ProTα resonances decreases, and those of the globular domain drop below the noise (Fig. 1f, g, Extended Data Fig. 3b). The large hydrodynamic radii of H1 and the complex (Extended Data Fig. 4a, b) support a large effective rotational correlation time as the origin of peak broadening, but a contribution from chemical exchange cannot be excluded. However, the globular domain is dispensable for complex formation (Fig. 2b).

High-affinity binding in spite of disorder

To quantify the strength of the interaction between H1 and ProTα, we used single-molecule Förster resonance energy transfer (FRET), which enables measurements over a very broad range of affinities down to the picomolar regime. By labelling two positions with a donor and an acceptor dye, distances and distance changes between or within the polypeptides can be determined by confocal fluorescence detection of molecules freely diffusing in solution[27,28]. ProTα labelled at positions 56 and 110 (ProTα$_{56/110}$; all labelled residues are cysteines) exhibits a mean transfer efficiency, ⟨*E*⟩, of 0.33 at near-physiological ionic strength (Fig. 2a, Extended Data Table 2), as expected for this IDP, which is highly expanded owing to its large negative net charge[13,29,30]. On addition of unlabelled H1, a population with higher ⟨*E*⟩ of 0.58 (that is, shorter average distance within the ProTα chain) emerges: binding to the positively charged H1 evidently leads to a compaction of ProTα by charge screening, analogous to the compaction on addition of salt[29]. The same behaviour is observed for doubly labelled H1 (Extended Data Table 2), which demonstrates mutual adaptation of the conformational ensembles. The resulting dissociation constant in the low picomolar range reveals an extremely strong interaction (Fig. 2b, Extended Data Table 2), consistent with the physiological role of ProTα as a linker histone chaperone[17] that competes with the tight binding of H1 to chromatin[31]. Measurements with other FRET dyes and label positions resulted in similar affinities (Extended Data Table 2), indicating that labelling has only a small effect on binding. The dominant contribution to the interaction with ProTα stems from the unstructured C-terminal part of H1, which in isolation still binds with picomolar affinity. The N-terminal half and the isolated globular domain of H1 also bind to ProTα, but with much lower affinity (Fig. 2b). At least four isolated globular domains can bind to one ProTα molecule at the same time, with modest chemical shift changes (Extended Data Fig. 1), suggesting the absence of a specific binding interface.

The large and opposite net charges of ProTα (−44) and H1 (+53) imply a strong electrostatic contribution to binding. Indeed, a mere doubling of the ionic strength from the physiological 165 mM to

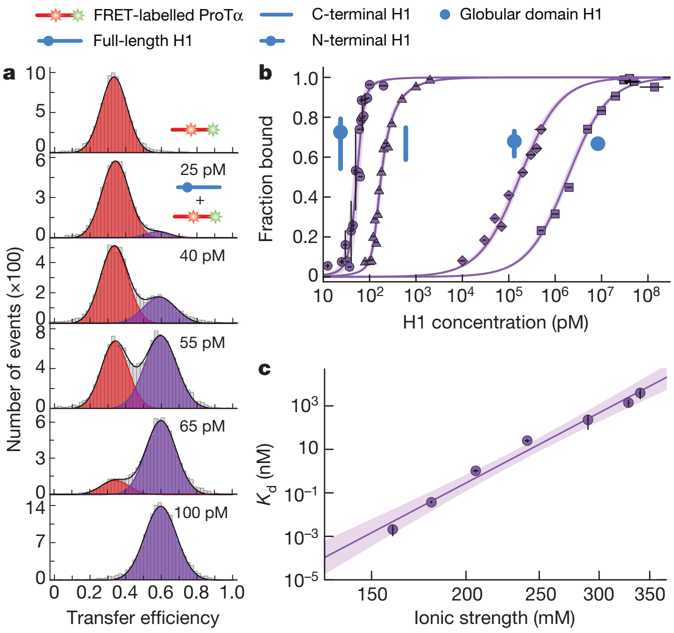


Figure 2 | ProTα and H1 form an electrostatically driven high-affinity complex. a, Single-molecule transfer efficiency histograms of ProTα$_{56/110}$ (FRET-labelled at positions 56 and 110) with varying concentrations (0 pM, 25 pM, 40 pM, 55 pM, 65 pM or 100 pM) of unlabelled H1 as indicated in the panels, fitted with an unbound peak (red) and a bound peak (purple). **b**, Binding isotherms based on transfer efficiency histograms for full-length H1 (circles, $K_d = 2.1^{+1.1}_{-0.8}$ pM), H1 N-terminal region (diamonds, $K_d = 173^{+29}_{-28}$ nM), H1 C-terminal region (triangles, $K_d = 40^{+6}_{-4}$ pM) and the globular domain of H1 (squares, $K_d = 1.9^{+0.3}_{-0.3}$ μM) at 165 mM ionic strength (see Extended Data Table 1 for details). **c**, K_d of H1–ProTα complex as a function of ionic strength with fit in terms of counterion release[50] (purple line) and 95% confidence interval (shaded). See Methods for details of statistics and data analysis.

340 mM reduces the affinity by six orders of magnitude (Fig. 2c). By extrapolation, a reduction of ionic strength to approximately 140 mM would take this interaction into the femtomolar range. From low picomolar to 100 μM protein concentrations, the stoichiometry from intermolecular FRET (Extended Data Fig. 4c), NMR chemical shift titrations (Extended Data Figs 2, 3) as well as the hydrodynamic radii measured with pulsed-field gradient NMR and two-focus fluorescence correlation spectroscopy (Extended Data Fig. 4a, b) all indicate the predominant formation of one-to-one dimers and the absence of large oligomers or coacervates[32]. However, in the presence of a large excess of one of the binding partners, we observe a decrease in FRET efficiencies that is indicative of the weak association of additional molecules with a dissociation constant (K_d) in the 10 to 100 μM range (Extended Data Fig. 4d, e), a propensity that is also observed in the simulations described later.

A highly dynamic complex

The lack of structure formation in the H1–ProTα complex implies great flexibility and a highly dynamic interconversion within a large ensemble of configurations and relative arrangements of the two IDPs. The presence of a broad and rapidly sampled distance distribution is supported by the analysis of fluorescence lifetimes[28,33,34] (Extended Data Fig. 5). Because fluctuations in distance cause fluctuations in the fluorescence intensity of donor and acceptor, the timescale of these long-range distance dynamics can be measured by single-molecule FRET combined with nanosecond fluorescence correlation spectroscopy[34,35]. For unfolded or disordered proteins, reconfiguration times (the relaxation times for distances within the chain) of approximately 20–200 ns have previously been observed[27]. ProTα is a particularly dynamic IDP, because of its highly expanded chain[13,29] and corresponding lack of impeding intramolecular interactions[36]; in isolation it yields reconfiguration times (τ_r) between 29^{+2}_{-2} ns and 78^{+15}_{-9} ns, depending on the chain segment probed[34,36] (Extended Data Table 2). H1 labelled at positions 113 and 194 reconfigures more slowly than ProTα ($\tau_r = 118^{+24}_{-14}$ ns), but is within the range previously observed for unfolded and disordered proteins[27,34].

Notably, these pronounced and rapid long-range dynamics are retained in the H1–ProTα complex; we observed values for τ_r of between 66^{+2}_{-2} ns and 191^{+22}_{-19} ns for 13 different labelling pairs throughout the dimer (Fig. 3a–d, Extended Data Table 2). The similarity between the τ_r values of H1 and ProTα when bound in the complex suggests a coupling of the dynamics of the two intertwining chains. The highly dynamic nature of the complex is further supported by NMR: the longitudinal (T_1) and transverse (T_2) ^{15}N relaxation times reflect rapid backbone dynamics in the pico- to nanosecond range for ProTα in both the free and the bound state (Fig. 3h, Extended Data Fig. 2). The increase in T_1/T_2 (Fig. 3h) and in the hydrodynamic radius (Extended Data Fig. 4), as well as the reduced peak intensities (Fig. 3f) are consistent with the increase in τ_r for ProTα observed by nanosecond fluorescence correlation spectroscopy in the complex (Fig. 3a), in which chain–chain interactions are expected to slow down both local and long-range dynamics.

Architecture of an unstructured protein complex

To develop a structural representation of the conformational ensemble of the H1–ProTα complex, we combined single-molecule FRET, NMR and molecular simulations. We first mapped the complex with single-molecule FRET by probing a total of 28 intra- and intermolecular distances with donor and acceptor dyes in specific positions (Figs 3i, 4a). The resulting intermolecular transfer efficiencies lack the pronounced patterns that would be expected if persistent site-specific interactions or chain alignment in a preferred register were present. The intermolecular transfer efficiencies are most sensitive to the labelling position on ProTα, with the highest efficiencies (that is, shortest average distances) for the central position at residue 56 (ProTα$_{56}$), intermediate efficiencies for ProTα$_{110}$ and lowest efficiencies (that is, longest average distances) for ProTα$_2$. These results indicate that the region of highest charge density of ProTα (Fig. 1b) most strongly attracts H1. The charge density along H1 is more uniform (Fig. 1a), as are the transfer efficiencies to ProTα, albeit with some decrease towards the termini (Fig. 3i).

On the basis of this information we sought to establish a molecular model of the H1–ProTα complex. Given the lack of structure formation and residue-specific interactions, the dominance of electrostatics and the size of the system, we used a simplified coarse-grained model in which each residue is represented by a single bead. Coulombic interactions between all charged residues are included explicitly, with a screening factor to account for an ionic strength of 165 mM. Other attractive interactions and excluded volume repulsion are captured using a short-range potential, with the radius of the residues determined from their volumes[37]. A structure-based potential[38] is used to describe the folded globular domain of H1. The transfer efficiencies computed from Langevin dynamics simulations can be matched to the measured values (Fig. 4a) via the single adjustable parameter in our model—namely, the contact energy of the short-range potential—which is the same for all residues (see Methods); explicitly including a representation of the chromophores in the simulations yielded very similar results (Fig. 4a). The resulting intra- and intermolecular distance distributions (Extended Data Fig. 6d) are smooth and unimodal, which is consistent with the absence of site-specific interactions and structure formation observed experimentally and attests to the convergence of the simulations. The good agreement between the transfer efficiencies observed in our experiments and those obtained from the simulation indicates that this simple model captures the essential properties of the structural ensemble. Considering its simplicity, the femtomolar affinity estimated from the model

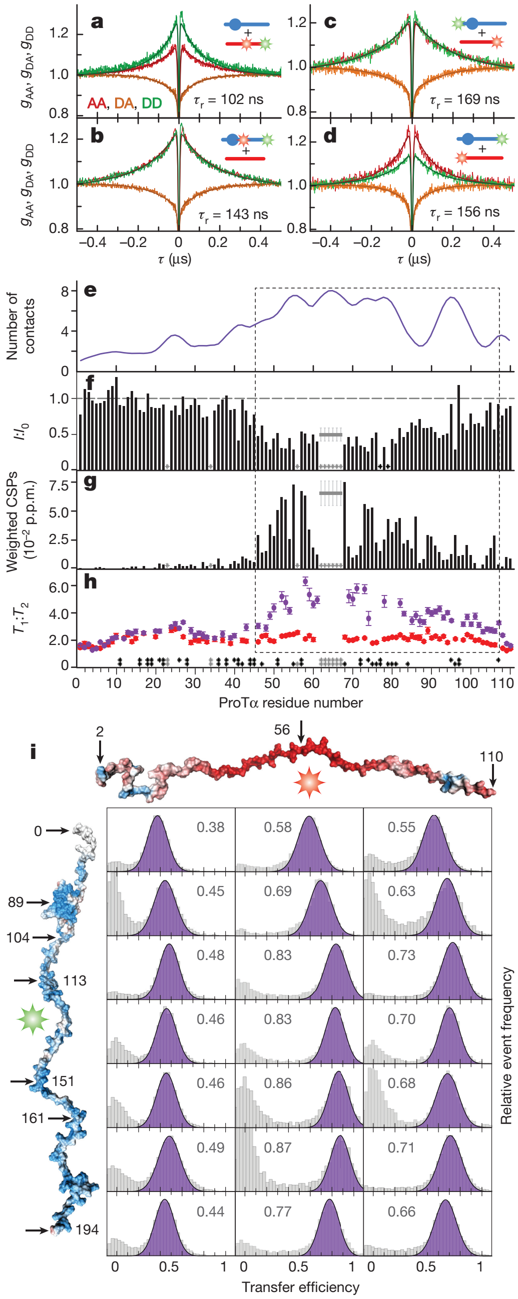


Figure 3 | Dynamics, interactions, and distances in the complex. **a**–**d**, Examples of nanosecond fluorescence correlation spectroscopy probing long-range dynamics, based on intra- and intermolecular FRET (see Extended Data Table 2 for details); curves are the averages of $n = 3$ independent measurements of acceptor–acceptor (g_{AA}), donor–acceptor (g_{DA}) and donor–donor (g_{DD}) correlation, performed with doubly labelled ProTα with unlabelled H1 (**a**), doubly labelled H1 with unlabelled ProTα (**b**) and singly labelled H1 and singly labelled ProTα (**c**, **d**). Pictograms as in key in Fig. 2. **e**, Average number of contacts of each ProTα residue with H1 based on the simulations (Fig. 4b). **f**, Ratios of NMR resonance intensities of ProTα in the presence (I) and absence (I_0) of H1. **g**, Weighted backbone amide chemical shift perturbations (CSPs) of ProTα induced by equimolar H1 binding (see Extended Data Fig. 2 for other stoichiometries); $n = 5$ repeats of this measurement yielded consistent results. In **f** and **g**, the grey horizontal lines represent the average of three unassigned but traceable Glu residues in the range of residues 62–67 with error bars from their s.d. (see Methods). **h**, Ratios of T_1 and T_2 ^{15}N relaxation times of ProTα in free (red) and bound (purple) states (see Extended Data Fig. 2 for details). Light grey stars, prolines and unassigned residues; black stars, resonance overlap and/or insufficient data quality. Circles are mean values from $n = 3$ consecutive measurements, errors are s.d. The dashed box in **e**–**h** indicates the sequence range with the largest changes. **i**, Single-molecule transfer efficiency histograms from intermolecular FRET experiments between different positions in acceptor-labelled (red star) ProTα and donor-labelled (green star) H1, fitted with a single peak (purple, mean transfer efficiency shown). The signal at $E \approx 0$ originates from molecules without active FRET acceptor. For further information on statistics, see Methods.

(Extended Data Fig. 5b) is remarkably consistent with the affinities that were observed experimentally near this ionic strength (165 mM). The affinity for a second molecule of H1 or ProTα to the complex is predicted to be orders of magnitude weaker than for the first molecule, consistent with the experimental results (Extended Data Figs 4d, e, 6b).

The intra- and intermolecular distance maps from the simulation (Fig. 4b) indicate that the interactions between ProTα and H1 are broadly distributed along their sequences, but also reflect the asymmetry in electrostatic attraction owing to the higher charge density of ProTα in its central and C-terminal regions (Figs 1b, 4a). The NMR results provide an independent experimental test of the model: the distribution of the average number of contacts made by the residues of ProTα based on the simulation (Fig. 3e) is notably similar to the distribution of changes in chemical shifts, peak intensities and T_1/T_2 ratios observed on binding (Fig. 3f–h). These changes occur across the same broad region between residues 46 and 106, encompassing the most acidic tracts of ProTα. Overlap within the Glu cluster of the NMR spectra prevents the quantitative analysis of some of the corresponding peaks, but the chemical shift and intensity perturbations in this cluster are similar to those observed in the rest of the region from residues 46 to 106 (Fig. 3f, g).

Further analysis of the simulated structural ensemble (see Supplementary Video 1) shows a lack of distinct conformational clusters (Extended Data Fig. 6a), which implies a continuous distribution of configurations. A projection of the simulation onto the first three principal components of the inter-residue distances (Extended Data Fig. 6c) reveals a highly heterogeneous ensemble of arrangements of the two entwining flexible chains (Fig. 4c). Given the rapid intramolecular dynamics and lack of structure in the complex, the activation barrier for binding is likely to be close to zero. Indeed, association of H1 and ProTα occurs at the diffusion limit, with a binding rate coefficient of $3.1 \pm 0.1 \times 10^9$ M^{-1} s^{-1} (Extended Data Fig. 7). The simulations support this mechanism, with a downhill free-energy surface for binding and attractive fly-casting[39] interactions enhanced by electrostatics[40] already emerging at a distance of approximately 22 nm, which is much greater than the sum of the hydrodynamic radii (Extended Data Fig. 6b).

Conclusions

Our results suggest that high-affinity complex formation between two oppositely charged IDPs is possible without the formation of structure or the need for folded domains. In contrast to the current paradigm for molecular recognition in biomolecular interactions, this type of highly dynamic complex does not require structurally defined binding sites or specific persistent interactions between individual residues. Instead, our findings are well-described as being the result of long-range electrostatic attraction between the two interpenetrating polypeptide chains, especially between their charge-rich regions. The exceedingly rapid interconversion of many different arrangements and configurations on the 100-ns timescale results in efficient averaging and essentially corresponds to a mean-field-type interaction[41,42] between all charges. This type of complex expands the known spectrum of protein–protein interactions. Although the H1–ProTα complex is extreme in the extent of its disorder for both binding partners, the possibility of this interaction mechanism is not entirely unexpected, given the prevalence of charged amino acids in many IDPs[2], the previous observation of disorder in IDPs interacting with folded proteins[3–7]

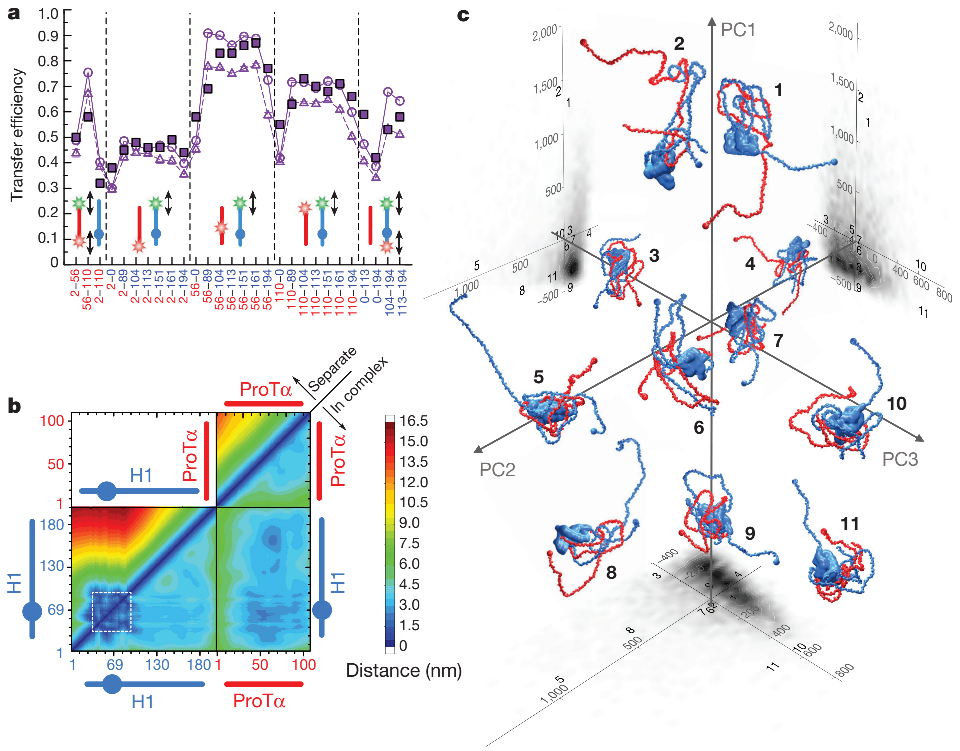


Figure 4 | Architecture of the complex from simulations. a, Comparison of experimental (filled squares) and simulated transfer efficiencies (empty symbols) in the H1–ProTα complex for the pairs of dye positions indicated on the *x* axis (red, position on ProTα; blue, position on H1). Triangles and circles indicate simulations with and without explicit chromophores, respectively. Pictograms as in key in Fig. 2. **b,** Intra- and intermolecular average distance maps of H1 and ProTα from the simulations, separately and in complex. The white dashed square indicates the globular domain (only surface-exposed residues shown, see Extended Data Table 1). **c**, Examples of configurations of H1 (blue) and ProTα (red) in the complex; N termini are indicated by small spheres. The structures are projected onto the first three principal components (PC1, PC2 and PC3) of the distance map, with projections of the full ensemble shown as grey scatter plots (units of Å, see also Extended Data Fig. 6 and Supplementary Video 1). Bold numbers indicate the positions of the structures in the principal component projections.

and the role of electrostatics in the formation of dynamic binding interfaces between folded proteins[43]. Moreover, the H1–ProTα interaction resembles polyelectrolyte complexes formed by charged synthetic polymers[42], even though the latter usually phase-separate into coacervates[42]. The absence of coacervation[32,42] or liquid–liquid phase separation[9] for ProTα and H1 at concentrations ranging from the picomolar to high micromolar regime may be due to the complementarity[44] of the two proteins in terms of effective length and opposite net charge, leading to optimal and mutually saturating electrostatic interactions, or to the lack of hydrophobic and aromatic side chains and cation–π interactions, which have previously been suggested to favour phase separation mediated by proteins (refs 32, 45 and R. M. Vernon *et al*., unpublished)

There are several functional implications of this mechanism for forming a high-affinity yet unstructured dynamic complex between two IDPs. Histone H1 is a key factor in chromatin condensation and transcriptional regulation[11], and ProTα acts as a chaperone of H1 that facilitates its displacement from and deposition onto chromatin[17]. ProTα thus needs to be able to compete with the very high affinity of the histone to chromatin[31]. However, high affinities between structured biomolecules are usually linked to exceedingly slow dissociation[40], which is incompatible with fast regulation. By contrast, the high affinity of the H1–ProTα complex is facilitated by its ultra-fast association, which allows dissociation on a biologically relevant timescale in spite of the high affinity required for function. Another consequence of polyelectrolyte interactions is the possibility of ternary complex formation[46], signs of which are detected here with a large excess of ProTα or H1 (Extended Data Figs 4d, e, 6b), resulting in mostly unexplored kinetic mechanisms that cannot be explained by competition via simple dissociation and re-association[47]. Finally, the flexibility within such unstructured complexes may facilitate access for enzymes that add posttranslational modifications; these modifications have key roles in the regulation of cellular processes, including those of H1. One example of this mechanism may be the interaction of the acidic domain of the oncogene SET with the lysine-rich C-terminal tail of p53, which is regulated by acetylation[48].

The behaviour we observe for ProTα and H1 might be surprisingly widespread, as highly charged protein sequences that could form such complexes are abundant in eukaryotes. In the human proteome alone, several hundred proteins that are predicted to be intrinsically disordered[49] contain contiguous stretches of at least 50 residues with a fractional net charge similar to that of H1 or ProTα. As the interaction of highly oppositely charged IDPs is unlikely to be very sequence-specific[18], achieving binding selectivity may be linked to other regulatory mechanisms such as cellular localization or synchronized expression during relevant stages of development or the cell cycle.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 August 2017; accepted 17 January 2018.

Published online 21 February 2018.

1. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19,** 31–38 (2009).
2. Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chem. Rev.* **114,** 6561–6588 (2014).
3. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33,** 2–8 (2008).
4. Baker, J. M. *et al.* CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nat. Struct. Mol. Biol.* **14,** 738–745 (2007).
5. Mittag, T. *et al.* Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl Acad. Sci. USA* **105,** 17772–17777 (2008).
6. Hendus-Altenburger, R. *et al.* The human Na$^+$/H$^+$ exchanger 1 is a membrane scaffold protein for extracellular signal-regulated kinase 2. *BMC Biol.* **14,** 31 (2016).
7. Milles, S. *et al.* Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell* **163,** 734–745 (2015).
8. Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic protein interaction networks and new structural paradigms in signaling. *Chem. Rev.* **116,** 6424–6462 (2016).
9. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18,** 285–298 (2017).
10. Robinson, P. J. & Rhodes, D. Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Curr. Opin. Struct. Biol.* **16,** 336–343 (2006).
11. Hergeth, S. P. & Schneider, R. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Rep.* **16,** 1439–1453 (2015).
12. Hansen, J. C., Lu, X., Ross, E. D. & Woody, R. W. Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* **281,** 1853–1856 (2006).
13. Gast, K. *et al.* Prothymosin α: a biologically active protein with random coil conformation. *Biochemistry* **34,** 13211–13218 (1995).

14. Uversky, V. N. *et al.* Natively unfolded human prothymosin α adopts partially folded collapsed conformation at acidic pH. *Biochemistry* **38,** 15009–15016 (1999).
15. Gómez-Márquez, J. & Rodríguez, P. Prothymosin α is a chromatin-remodelling protein in mammalian cells. *Biochem. J.* **333,** 1–3 (1998).
16. Mosoian, A. Intracellular and extracellular cytokine-like functions of prothymosin α: implications for the development of immunotherapies. *Future Med. Chem.* **3,** 1199–1208 (2011).
17. George, E. M. & Brown, D. T. Prothymosin α is a component of a linker histone chaperone. *FEBS Lett.* **584,** 2833–2836 (2010).
18. Papamarcaki, T. & Tsolas, O. Prothymosin α binds to histone H1 *in vitro*. *FEBS Lett.* **345,** 71–75 (1994).
19. Barbero, J. L., Franco, L., Montero, F. & Morán, F. Structural studies on histones H1. Circular dichroism and difference spectroscopy of the histones H1 and their trypsin-resistant cores from calf thymus and from the fruit fly *Ceratitis capitata*. *Biochemistry* **19,** 4080–4087 (1980).
20. Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362,** 219–223 (1993).
21. Yi, S., Brickenden, A. & Choy, W. Y. A new protocol for high-yield purification of recombinant human prothymosin α expressed in *Escherichia coli* for NMR studies. *Protein Expr. Purif.* **57,** 1–8 (2008).
22. Khan, H. *et al.* Fuzzy complex formation between the intrinsically disordered prothymosin α and the Kelch domain of Keap1 involved in the oxidative stress response. *J. Mol. Biol.* **425,** 1011–1027 (2013).
23. Zhou, B. R. *et al.* Structural insights into the histone H1–nucleosome complex. *Proc. Natl Acad. Sci. USA* **110,** 19390–19395 (2013).
24. Zarbock, J., Clore, G. M. & Gronenborn, A. M. Nuclear magnetic resonance study of the globular domain of chicken histone H5: resonance assignment and secondary structure. *Proc. Natl Acad. Sci. USA* **83,** 7628–7632 (1986).
25. Kjaergaard, M., Brander, S. & Poulsen, F. M. Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *J. Biomol. NMR* **49,** 139–149 (2011).
26. Bae, S. H., Dyson, H. J. & Wright, P. E. Prediction of the rotational tumbling time for proteins with disordered segments. *J. Am. Chem. Soc.* **131,** 6814–6821 (2009).
27. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.* **45,** 207–231 (2016).
28. Sisamakis, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate single-molecule FRET studies using multiparameter fluorescence detection. *Methods Enzymol.* **475,** 455–514 (2010).
29. Müller-Späth, S. *et al.* Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl Acad. Sci. USA* **107,** 14609–14614 (2010).
30. Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl Acad. Sci. USA* **109,** 16155–16160 (2012).
31. White, A. E., Hieb, A. R. & Luger, K. A quantitative investigation of linker histone interactions with nucleosomes and chromatin. *Sci. Rep.* **6,** 19122 (2016).
32. Pak, C. W. *et al.* Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell* **63,** 72–85 (2016).
33. Chung, H. S., Louis, J. M. & Gopich, I. V. Analysis of fluorescence lifetime and energy transfer efficiency in single-molecule photon trajectories of fast-folding proteins. *J. Phys. Chem. B* **120,** 680–699 (2016).
34. Soranno, A. *et al.* Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl Acad. Sci. USA* **109,** 17800–17806 (2012).
35. Nettels, D., Gopich, I. V., Hoffmann, A. & Schuler, B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl Acad. Sci. USA* **104,** 2655–2660 (2007).
36. Soranno, A. *et al.* Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl Acad. Sci. USA* **114,** E1833–E1839 (2017).
37. Creigthon, T. E. *Proteins: Structures and Molecular Properties* 2nd edn (W. H. Freeman and Co., 1993).
38. Karanicolas, J. & Brooks, C. L. III. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11,** 2351–2361 (2002).
39. Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl Acad. Sci. USA* **97,** 8868–8873 (2000).
40. Schreiber, G., Haran, G. & Zhou, H. X. Fundamental aspects of protein–protein association kinetics. *Chem. Rev.* **109,** 839–860 (2009).
41. Borg, M. *et al.* Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl Acad. Sci. USA* **104,** 9650–9655 (2007).
42. Srivastava, S. & Tirrell, M. V. in *Advances in Chemical Physics* (eds Rice, S. A. & Dinner, A. R.) Ch. 7, 499–544 (John Wiley & Sons, 2016).
43. Ahmad, A.*et al.* Heat shock protein 70 kDa chaperone/DnaJ cochaperone complex employs an unusual dynamic interface. *Proc. Natl Acad. Sci. USA* **108,** 18966–18971 (2011).
44. Freeman Rosenzweig, E. S. *et al.* The eukaryotic CO_2-concentrating organelle is liquid-like and exhibits dynamic reorganization. *Cell* **171,** 148–162.e119 (2017).
45. Nott, T. J. *et al.* Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57,** 936–947 (2015).
46. Peng, B. & Muthukumar, M. Modeling competitive substitution in a polyelectrolyte complex. *J. Chem. Phys.* **143,** 243133 (2015).
47. Berlow, R. B., Dyson, H. J. & Wright, P. E. Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature* **543,** 447–451 (2017).
48. Wang, D.*et al.* Acetylation-regulated interaction between p53 and SET reveals a widespread regulatory mode. *Nature* **538,** 118–122 (2016).
49. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434 (2005).
50. Record, M. T. Jr, Anderson, C. F. & Lohman, T. M. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Q. Rev. Biophys.* **11,** 103–178 (1978).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. A. Sjørup and J. H. Martinsen for purification assistance, I. König for a sample of wild-type ProTα, A. Prestel and M. B. Kunze for NMR advice, E. Holmstrom for help with data analysis, D. Mercadante for assistance with electrostatics calculations, R. Sobrino for help with experiments in the early stages of the project, J. Forman-Kay, M. Blackledge, R. Pappu and A. Holehouse for discussions and the Functional Genomics Center Zurich for performing mass spectrometry. This work was supported by the Swiss National Science Foundation (B.S.), the Danish Council for Independent Research (# 4181-00344, B.B.K.), the Novo Nordisk Foundation (B.B.K. and P.O.H.), the Carlsberg Foundation (P.O.H.), and the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (R.B.B.). This work used the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Author Contributions A.B., M.B.B., K.B., B.B.K., R.B.B. and B.S. designed research; M.B.B., A.B., V.M.K. and A.Sot. produced and labelled fluorescent protein variants; A.B. and M.B.B. performed single-molecule experiments; A.B., M.B.B., A.Sor. and D.N. analysed single-molecule data; D.N. developed single-molecule instrumentation and data analysis tools; A.Sot. and A.B. carried out stopped-flow measurements, A.B., M.B.B., K.J.B. and A.Sot. established experimental conditions for single-molecule measurements; C.B.F. and P.O.H. produced protein samples for NMR; K.B. and C.B.F. performed and analysed NMR measurements; A.Sor. carried out the bioinformatics analysis; R.B.B. conducted and analysed simulations; A.B., B.B.K. and C.B.F. carried out circular dichroism experiments; B.B.K., R.B.B. and B.S. supervised research; B.S., A.B., R.B.B., B.B.K. and K.B. wrote the paper with help from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to B.S. (schuler@bioc.uzh.ch), A.B. (a.borgia@bioc.uzh.ch), R.B.B. (robert.best2@nih.gov) or B.B.K. (bbk@bio.ku.dk).

Reviewer Information *Nature* thanks E. Zuiderweg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample sizes. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein preparation. ProTα was prepared and labelled as previously described[29]. For a complete list of all protein variants, labelling positions and fluorophores used for single-molecule FRET, see Extended Data Table 1. The correct molecular mass of all protein variants and labelled constructs was confirmed by mass spectrometry.

For experiments using wild-type human linker histone H1.0 (H1), recombinant protein was used (New England Biolabs, cat.# M2501S). For the production of labelled H1 variants and wild-type H1 for NMR, the cDNA of the human *H1F0* gene (UniProt P07305) was cloned into a modified version of the pRSET vector[51]. In this plasmid, the N-terminal hexahistidine tag and thrombin cleavage site were removed and replaced by a C-terminal hexahistidine tag and thrombin cleavage site (with sequence GGPRGSRGSHHHHHH) to enable purification of H1 free of degradation products. Cysteine mutations were introduced using site-directed mutagenesis, to enable labelling with fluorescent dyes using maleimide coupling (see Extended Data Table 1 for a complete list of variants). All H1 variants were expressed in *Escherichia coli* C41 cells and terrific broth medium at 37 °C, induced with 0.5 mM isopropylthiogalactopyranoside (IPTG) at an OD_{600} of ~0.6, and grown for 3 further hours. Cell pellets were collected and resuspended in denaturing buffer (6 M guanidinium chloride (GdmCl) in phosphate-buffered saline (PBS, 10 mM sodium phosphate pH 7.4, 137 mM NaCl, 2.7 mM KCl), the soluble fraction was collected and applied to a Ni-IDA resin (ABT Beads) in batch. The resin was washed twice with 5 resin volumes of denaturing buffer including 25 mM imidazole, three times with 5 resin volumes of PBS including 25 mM imidazole, and the protein was eluted with PBS including 250 or 500 mM imidazole. The protein was dialysed against PBS, filtered and its hexahistidine tag cleaved off with 5 U of thrombin (Serva) per milligram of H1, for 2 h at room temperature. To remove uncleaved protein and the tag, the mixture was run through a HisTrap HP 5-ml column (GE Healthcare) in PBS including 25 mM imidazole. H1 was further purified using a Mono S ion exchange chromatography column (GE Healthcare), washed with 20 mM Tris (pH 8.0) including 200 mM NaCl, and eluted in 20 mM Tris (pH 8.0) buffer with a gradient from 200 mM to 1M NaCl. Finally, samples for labelling were reduced with 20 mM dithiothreitol and purified by reversed-phase high-performance liquid chromatography (HPLC) with a Reprosil Gold C4 column with a gradient from 5% acetonitrile and 0.1% trifluoroacetic acid in aqueous solution to 100% acetonitrile. H1-containing fractions were lyophilized and resuspended in degassed 6 M GdmCl, 50 mM sodium phosphate buffer (pH 7.0). For double labelling, both dyes (dissolved in dimethylsulfoxide) were added to the protein in a 1:1:1 molar ratio; for single labelling, dye was added at a 0.7:1 molar ratio of dye to protein. Reactions were incubated at room temperature for 2 h, and stopped by adding 20 mM dithiothreitol. Products were purified by reversed-phase HPLC, and the correct mass of all labelled proteins confirmed by mass spectrometry (see Extended Data Fig. 8 for an example). Lyophilized labelled protein was dissolved in 8 M GdmCl and stored at −80 °C.

For NMR experiments, H1, ProTα and the H1 globular domain (Extended Data Table 1) were either produced in unlabelled form by growing cells in LB medium, uniformly labelled with ^{15}N by growing cells in M9 minimal medium containing $^{15}NH_4Cl$ as the sole source of nitrogen, or uniformly labelled with ^{15}N and ^{13}C by growing cells in M9 minimal medium containing $^{15}NH_4Cl$ as the sole source of nitrogen and $^{13}C_6$-glucose as the sole source of carbon, as previously described[6] and then purified essentially as explained earlier. The H1 globular domain was expressed as a GST-fusion protein with a TEV protease site, and purified on a glutathione sepharose 4 fast-flow column (GE Healthcare). The column was washed with 10 column volumes (CV) of PBS and the tagged protein eluted with 5 CV of elution buffer (50 mM Tris-HCl pH 8.0, 10 mM reduced glutathione). All fractions containing GST–H1 globular domain were pooled and cleaved with TEV protease (100 μl of 0.5 mg ml^{-1} stock solution) overnight, and subsequently applied to a HiTrap SP FF 5 mL (GE Healthcare) with 50 mM sodium phosphate pH 9.0 and eluted with 50 mM sodium phosphate pH 9.0, 1 M NaCl over 25 CV. The protein-containing fractions were applied to a Superdex 75 10/300 (GE Healthcare) in TBS buffer (10 mM Tris, 157 mM KCl, 0.1 mM EDTA, pH 7.4) and further concentrated using an Amicon Ultra-15 centrifugal filter device (Millipore) with a molecular weight cutoff of 3 kDa. Protein concentrations of H1 and H1 globular domain were determined by UV absorbance, and the concentration of ProTα was determined by BCA assay (Thermo Fisher Scientific).

NMR spectroscopy. To minimize amide exchange, all NMR spectra were acquired at 283 K, unless otherwise specified, on a Varian INOVA 800-MHz (^1H) spectrometer with a room temperature probe or Bruker AVANCE III 600- or 750-MHz (^1H) spectrometers equipped with cryogenic probes. Free induction decays were transformed and visualized in NMRPipe[52] or Topspin (Bruker Biospin) and analysed using CcpNmr Analysis software[53]. Assignments of backbone nuclei of ^{13}C–^{15}N-labelled ProTα in the unbound state (0.1 mM ^{13}C–^{15}N-labelled ProTα, TBS buffer, 10% D_2O (v/v), 0.7 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS)) and at sub-saturating concentration (1:0.8 molar ratio) of H1 (0.1 mM ^{13}C–^{15}N-labelled ProTα, 0.08 mM H1, TBS buffer, 10% D_2O (v/v), 0.7 mM DSS) were performed manually from the analysis of ^1H–^{15}N HSQC, HNCACB, CBCA(CO)NH, HN(CO)CA, HNCO and HN(CA)NNH spectra acquired with non-uniform sampling[54] using standard pulse sequences. At saturating concentrations of H1, backbone resonances of ProTα became too weak for successful assignments. Proton chemical shifts were referenced internally to DSS at 0.00 p.p.m., with heteronuclei referenced by relative gyromagnetic ratios. The content of transient structure in ProTα was evaluated for each state from secondary Cα-chemical shifts assigned in the free form and at 80% saturation of H1 using a random coil reference set for IDPs[25]. In both states, three transient marginally populated α-helices were identified: residues Ser9–Glu19 (~10% populated), Ala82–Thr86 (~13% populated) and Val99–Lys102 (~18% populated). The populations of the transient α-helices were estimated from the average SCS value of the residues of the transient helices divided by 2.8 p.p.m. ($SCS_{C\alpha}$ value expected for a fully populated α-helix)[55] and were very similar in the free and bound states. ^1H–^{15}N HSQC spectra of ^{15}N-labelled H1 (40 μM) were recorded in the absence and presence of ProTα (40 μM). ^1H–^{13}C HSQC and/or ^1H–^{15}N HSQC spectra were acquired on four different sequential titrations: addition of up to 44 μM H1 to 11 μM ^{15}N-labelled ProTα; addition of up to 140 μM H1 globular domain to 20 μM ^{15}N-labelled ProTα; addition of up to 400 μM ProTα to 100 μM ^{13}C–^{15}N-labelled H1–Gly-Ser-6 × His; and addition of up to 700 μM ProTα to 100 μM ^{13}C–^{15}N-labelled H1 globular domain. Before each titration, the proteins were concentrated and dialysed in the same beaker. Subsequently, the solution of labelled protein was split equally into two samples, to one of which the unlabelled titrant was added at the maximum concentration, and to the other the same volume of dialysis buffer. After acquisition of NMR spectra on the two samples, they were used to obtain titration points between the end points by sequentially mixing the sample of the complex into the free protein. All NMR titration data were recorded in TBS buffer, 10% D_2O (v/v), 0.7 mM DSS. Binding-induced weighted CSPs were calculated as the weighted Euclidean distance between the peaks using $|\gamma_N|/|\gamma_H| = 0.154$. Owing to extensive resonance overlap of H^N, N, Cα and Cβ resonances in the 2D and 3D NMR spectra, assignments of backbone nuclei were not possible for the Glu repeat region from Glu62–Glu67. Nonetheless, spin systems displaying resonances consistent with Glu residues with Glu neighbours could be identified, and by exclusion were assigned to be part of the Glu62–Glu67 Glu repeat. For three of these systems, amide backbone peaks could be confidently tracked in the titration of ^{15}N-labelled ProTα with H1. The intensity ratios and weighted CSPs of the three Glu amide backbone peaks on addition of equimolar H1 were calculated and the average value used to represent the Glu repeat region in Fig. 3f, g. Peaks from the remaining Glu residues were present in the spectra of both free and bound states of ProTα, but could not be followed unambiguously during the titrations.

The hydrodynamic radii (R_H) of ProTα alone and at saturating concentrations of H1 or H1 globular domain were determined from a series of ^1H–^{15}N HSQC spectra with preceding pulse-field gradient stimulated-echo longitudinal encode–decode diffusion filter[56] and with the gradient strength increasing linearly from 0.963 to 47.2 G cm^{-1}. To determine the diffusion coefficients (D) the decay curves of the amide peaks were plotted against the gradient strength and fitted in Dynamics Center (Bruker) using $I = I_0\exp(-Dx^2\gamma^2\delta^2(\Delta - \delta/3) \times 10^4)$, in which I is the intensity of the NMR signal at the respective gradient strength, I_0 the intensity without applied gradient, x the gradient strength in G cm^{-1}, $\gamma = 26752$ rad Gs^{-1}, $\delta = 3$ ms, $\Delta = 250$ ms. R_H was calculated from the diffusion coefficient using the Stokes–Einstein relation, $R_H = k_BT/(6\pi\eta D)$, with η being the viscosity of water at 283 K.

T_1 and T_2 ^{15}N relaxation times were determined from 2× two series of ^1H–^{15}N HSQC spectra with varying relaxation delays using the pulse sequence of reference[57], and using pulsed-field gradients to suppress solvent resonances. The series were recorded on free ^{15}N-ProTα and on ^{15}N-ProTα with saturating concentrations of H1 at 800 MHz (^1H), using eight (10 ms, 100 ms, 300 ms, 500 ms, 700 ms, 1,100 ms, 1,300 ms and 1,500 ms) and seven (50 ms, 90 ms, 130 ms, 190 ms, 230 ms, 390 ms and 490 ms) different relaxation delays for T_1 and T_2, respectively, plus triplicate measurements. The relaxation decays were fitted to single exponentials and relaxation times determined using CcpNmr Analysis software[53].

Single-molecule fluorescence spectroscopy. Single-molecule measurements were performed using either a custom-built confocal instrument[58] or a MicroTime 200, both equipped with a HydraHarp 400 counting module (PicoQuant). The donor dye was excited with light from a 485-nm diode laser (LDH-D-C-485, PicoQuant) at an average power of 100 μW at the sample. The laser was operated in continuous-wave mode or in pulsed mode with alternating excitation of the dyes, achieved using pulsed interleaved excitation[59]. The wavelength range used for acceptor excitation was selected with a z582/15 band pass filter (Chroma) from the emission of a supercontinuum laser (EXW-12 SuperK Extreme, NKT Photonics) driven at 20 MHz, which triggers interleaved pulses from the 485-nm diode laser used for donor excitation. Emitted photons were collected by the microscope objective (Olympus UplanApo 60×/1.20 W), focused onto a 100-μm pinhole, and then separated into four channels with a polarizing beam splitter and two dichroic mirrors (585DCXR, Chroma). Emission was additionally filtered by bandpass filters (ET525/50M and HQ650/100, Chroma) before being focused onto one of four single-photon avalanche detectors (Optoelectronics SPCM AQR-15, PerkinElmer or τ-SPADs, PicoQuant).

FRET efficiency histograms of doubly labelled ProTα and H1 were acquired on samples with concentrations of labelled protein between 10 and 100 pM. For intermolecular measurements, up to 500 pM of acceptor-labelled protein were used to ensure saturation of binding. Measurements were performed in TBS buffer (165 mM ionic strength) or in an analogous buffer with higher ionic strength (adjusted by increasing the KCl concentration, as noted), in the presence of 140 mM β-mercaptoethanol (Sigma-Aldrich) for photoprotection[60] and 0.01% Tween 20 (Pierce) to minimize surface adhesion[61]. To avoid the pronounced interaction of H1 with glass surfaces, more-inert polymer sample chambers (μ-Slide, ibidi) were used throughout. Transfer efficiencies were obtained from $E = n_A/(n_A + n_D)$, in which n_D and n_A are the numbers of donor and acceptor photons, respectively, in each burst, corrected for background, channel crosstalk, acceptor direct excitation, differences in quantum yields of the dyes and detection efficiencies[61]. Even in cases in which pulsed interleaved excitation was insufficient to completely eliminate the donor-only contribution to the signal (Fig. 3i), the population at zero transfer efficiency was sufficiently well separated from the FRET population that the reliability of the transfer efficiencies was not affected. Fluorescence anisotropy values were determined for all labelling positions using polarization-sensitive detection in the single-molecule instrument[28,62], and were between 0.04 and 0.16 for the monomeric proteins, and between 0.08 and 0.22 in the complex, indicating sufficiently rapid orientational averaging of the fluorophores to justify the approximation $\kappa^2 \approx 2/3$ used in Förster theory[63].

The low fluorescence anisotropy values, the consistency of the FRET and NMR results, and the self-consistency of a large number of labelling positions suggest that the fluorophores do not entail a severe perturbation of the interaction between ProTα and H1. However, to assess the effect of fluorophore labelling in more detail, we tested how different dye pairs and labelling positions influence the affinity between ProTα and H1 and the inferred inter-dye distances (Extended Data Table 2). In view of the high net charge of the proteins, alternative fluorophores with a net charge different from Alexa 488 and 594 (both net charge −2) were chosen: Cy3B (GE Healthcare Life Sciences; zwitterionic with zero net charge), Abberior STAR 635 (Abberior GmbH; zwitterionic with zero net charge), and Atto550 and Atto647N (ATTO-TEC; both net charge +1). The K_d values for the respective binding partner were between 1.0 nM and 3.5 nM (at 205 mM ionic strength to simplify quantification) for all labelling positions and dye pairs, corresponding to an energetic perturbation of binding by at most ~1 k_BT. To test for the effect of the fluorophores on the inferred distances, we recorded single-molecule transfer efficiency histograms of ProTα labelled at positions 56 and 110 (ProT$\alpha_{56/110}$) with Cy3B/Abberior STAR 635 and Atto550/Atto647N, and of H1 labelled at positions 104 and 194 (H$1_{104/194}$) with Cy3B/Abberior STAR 635, both with and without the respective unlabelled binding partner present. The resulting transfer efficiency values yielded root mean square interdye distances consistent with those inferred from measurements with Alexa 488/594 (assuming a Gaussian chain distribution of inter-dye distances[29] and an experimental uncertainty of ± 0.05 for the transfer efficiency due to instrument calibration for the different dye pairs).

Fluorescence lifetime analysis. The comparison of ratiometric transfer efficiencies with the mean fluorescence lifetimes of donor and acceptor provides a further diagnostic for the presence of a broad distance distribution rapidly sampled during the time of a fluorescence burst[28,33,34]. Average lifetimes were estimated by using the mean donor ($\langle t_D \rangle$) and acceptor ($\langle t_A \rangle$) arrival times of the respective photons in a burst relative to the exciting laser pulse, and were combined with transfer efficiencies in a two-dimensional plot (Extended Data Fig. 5), in which $\tau_D^D/\tau_D^0 = \langle t_D \rangle/\tau_D^0$ and $\tau_D^A/\tau_D^0 = (\langle t_A \rangle - \tau_A^0)/\tau_D^0$ were calculated for each burst. In these equations, τ_D^0 is the intrinsic donor lifetime in the absence of the acceptor, and τ_A^0 is the intrinsic acceptor lifetime. For a single, fixed inter-dye distance (and thus transfer efficiency, E), one finds $\tau_D^D/\tau_D^0 = \tau_D^A/\tau_D^0 = 1 - E$, as illustrated by the diagonal line in Extended Data Fig. 5.

Nanosecond fluorescence correlation spectroscopy. Data for nanosecond fluorescence correlation spectroscopy were acquired at a concentration of ~100 pM of the protein carrying the donor (or both donor and acceptor) and an excess of the partner (either unlabelled or acceptor-labelled) to saturate binding. Donor and acceptor fluorescence emission (on continuous-wave excitation at 485 nm) from the subpopulation corresponding to the H1–ProTα complex in a transfer efficiency histogram was correlated with a binning time of 1 ns. To avoid the effects of detector dead times and after-pulsing on the correlation functions, the signal was recorded with two detectors each for donor and acceptor and cross-correlated between detectors[34,35]. Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were computed from the measurements and analysed as previously described [34,64]. In brief, auto- and cross-correlation curves were fitted over a time window of 2.5 μs with

$$g_{ij}(\tau) = 1 + \frac{1}{N}(1 - c_{ab}e^{-|\tau|/\tau_{ab}})(1 + c_{cd}e^{-|\tau|/\tau_{cd}})(1 + c_T e^{-|\tau|/\tau_T}) \text{ and } i, j = A, D$$

in which i and j correspond to donor or acceptor fluorescence emission; N is the effective mean number of molecules in the confocal volume; c_{ab}, τ_{ab}, c_{cd} and τ_{cd} are the amplitudes and time constants of photon antibunching and chain dynamics, respectively; and c_T and τ_T refer to the triplet blinking component on the microsecond timescale. Distance dynamics result in a characteristic pattern of the correlation functions based on donor and acceptor emission, with a positive amplitude in the autocorrelations ($c_{cd} > 0$) and a negative amplitude in the cross-correlation ($c_{cd} < 0$), but with identical decay times. All three correlation curves were thus fitted globally with the same values of τ_{cd}. Independent values of c_{cd}, c_{ab}, τ_{ab}, τ_T and c_T were used as free-fit parameters for each correlation curve. τ_{cd} was converted to the reconfiguration time of the chain, τ_r, as previously desc$_r$ibed[64], by assuming that chain dynamics can be modelled as a diffusive process in the potential of mean force derived from the sampled inter-dye distance distribution $P(r)$[35,64]. In light of the good agreement between the transfer efficiencies observed experimentally and in the simulations, we employed the $P(r)$ distributions obtained from the simulations for the respective pairs of labelling sites (intra- or intermolecular). This conversion does not entail a large change in timescale, and τ_{cd} and τ_r differ by less than 20% in all cases investigated here, depending on the average distance relative to the Förster radius[64]. The correlation functions shown in Fig. 3a–d were normalized to 1 at their respective values at 0.5 μs to facilitate direct comparison.

Two-focus fluorescence correlation spectroscopy. Two-focus fluorescence correlation spectroscopy measurements of Alexa 594-labelled ProTα were performed at 295 K on a MicroTime 200 confocal microscope equipped with a differential interference contrast prism. Alexa 594 was excited alternatingly with two orthogonally polarized laser beams: one beam with a wavelength range centred at 582 nm, selected with a z582/15 band pass filter (Chroma) from the emission of a supercontinuum fibre laser (EXW-12 SuperK Extreme, NKT Photonics) driven at 20 MHz, triggers (interleaved) pulses from a second supercontinuum laser with wavelength-selected output at 585 ± 3 nm (Solea, PicoQuant), with a combined repetition rate of 40 MHz and a power of 15 μW per laser at the sample. The distance between the two foci, δ, was calibrated as previously described on the basis of sample standards quantified under identical conditions using dynamic light scattering[30,65], yielding a δ of 490 ± 15 nm at $\lambda_{ex} = 585$ nm, corresponding to a systematic error of 3% of the calculated value of the hydrodynamic radius R_H. The concentration of labelled protein used in these experiments was ~4 nM in TBS buffer in the presence of 140 mM β-mercaptoethanol and 0.01% Tween 20. Translational diffusion coefficients were obtained from fits of the correlation functions[66] and converted to R_H using the Stokes–Einstein equation.

Analysis of binding isotherms. At ionic strengths of 200 mM and above, binding titrations of ProTα and H1 were hyperbolic and could be described well with a Langmuir-type isotherm, valid when the ligand concentration is sufficiently large compared to the analyte concentration. For example, with H1 as the ligand and ProTα as the analyte,

$$\frac{c_{\mathrm{H1-ProT\alpha}}}{c_{\mathrm{ProT\alpha}}^{\mathrm{tot}}} = \frac{c_{\mathrm{H1}}^{\mathrm{tot}}}{K_d + c_{\mathrm{H1}}^{\mathrm{tot}}},$$

in which the subscripts to c indicate the concentration of the species (that is, c_x for species x), and c_x^{tot} the total concentration of x. However, below an ionic strength of about 200 mM the affinity of H1 for the surface of the sample chambers in which the measurements were performed was so high that the surface of the chamber noticeably competed with H1 binding by ProTα; the polymeric sample chambers we used already exhibit much lower affinity for H1 than glass surfaces, which are negatively charged. This results in a decrease in the effective H1 bulk concentration available for binding to ProTα that leads to a shift of the apparent midpoint of the titration to higher H1 concentrations and to a distortion of the curve to a non-hyperbolic shape. To account for this effect, we need to take into account two coupled equilibria, one for the adsorption of H1 to surface binding sites, S, and one for H1 binding to ProTα:

$$\mathrm{H1} + S \rightleftarrows \mathrm{H1S} \text{ and } \mathrm{H1} + \mathrm{ProT}\alpha \rightleftarrows \mathrm{H1{-}ProT}\alpha,$$

with the dissociation constants

$$K_\mathrm{d}^\mathrm{H1S} = \frac{c_\mathrm{H1} \times \Gamma_\mathrm{S}}{\Gamma_\mathrm{SH1}} \tag{1}$$

and

$$K_\mathrm{d}^{\mathrm{H1-ProT}\alpha} = \frac{c_\mathrm{H1} \times c_{\mathrm{ProT}\alpha}}{c_{\mathrm{H1-ProT}\alpha}} \tag{2}$$

in which c_H1, $c_{\mathrm{ProT}\alpha}$, and $c_{\mathrm{H1-ProT}\alpha}$ are the bulk concentrations of free H1, free ProTα and complex, respectively. Γ_S and Γ_SH1 are the surface concentrations (that is, binding sites per area) of free binding sites and of binding sites occupied by H1, respectively. The resulting corresponding total concentrations are given by three equations:

$$c_\mathrm{H1}^\mathrm{tot} = c_\mathrm{H1} + c_{\mathrm{H1-ProT}\alpha} + \alpha\ \Gamma_\mathrm{SH1} \tag{3}$$

$$c_{\mathrm{ProT}\alpha}^\mathrm{tot} = c_{\mathrm{ProT}\alpha} + c_{\mathrm{H1-ProT}\alpha} \tag{4}$$

and

$$\Gamma_{\mathrm{S\ tot}} = \Gamma_\mathrm{S} + \Gamma_\mathrm{SH1} \tag{5}$$

In these equations, α is the surface-to-volume ratio of the sample well. Equations (1) to (5) were solved for the fraction of H1-bound ProTα using Mathematica (Wolfram Research) and the solution used to fit the titrations with full-length H1 and the H1 C-terminal fragment at 165 mM ionic strength (Fig. 2b) and full-length H1 at 185 mM ionic strength (Fig. 2c), with the adjustable parameters $K_\mathrm{d}^{\mathrm{H1-ProT}\alpha}$ and $K_\mathrm{d}^\mathrm{H1S}$, and with the product $\alpha \times \Gamma_\mathrm{SH1}$; $c_{\mathrm{ProT}\alpha}^\mathrm{tot}$ was fixed to the known value. The vertical error bars in Fig. 2b were estimated from five independent measurements. The horizontal error bars represent the pipetting errors estimated for the applied sequences of dilution steps. We obtained upper and lower bounds for the binding isotherms by taking into account the uncertainty of the ProTα concentration; these bounds are displayed as shaded bands in Fig. 2b. The resulting K_d for the full-length proteins at 165 mM ionic strength follows the trend expected from the measurements at higher ionic strength (Fig. 2c), validating the analysis. The weak association of additional monomers at high micromolar excess of binding partner was ignored in this analysis because it occurs in a different concentration regime. The dependence of the K_d on ion activity, a, (Fig. 2c) was analysed using a previously developed formalism[50], according to the approximation: $d\ln(K_d)/d\ln(a) \approx -\Delta\mathrm{n} = 18 \pm 1$ (standard error of the fit), in which $-\Delta n$ corresponds to the number of anionic and cationic counter ions released upon association of the two proteins, and the ion activity was approximated by the ionic strength.

Circular dichroism spectroscopy. Far-UV circular dichroism measurements were carried out on a Jasco J-810 spectropolarimeter, using a 1-mm path length quartz cuvette. Wild-type H1 and ProTα_{56} samples were measured at a concentration of 5 μM in TBS and 5 mM β-mercaptoethanol at 20 °C. A total of 20–60 spectra per sample were recorded between 250 and 195 nm with 1-nm step size, averaged and a buffer spectrum was subtracted. The far-UV circular dichroism spectrum of the H1 globular domain was recorded at 283 K from 260 to 198 nm with a scan speed of 20 nm/min, 10 accumulations and a response time of 2 s at a protein concentration of 10 μM in TBS, and the buffer spectrum was subtracted. To assess the thermal stability of the H1 globular domain, thermal unfolding was monitored at 222 nm from 283 to 378 K in increments of 1 K per minute. The ellipticity as a function of temperature was fitted with $\theta(T) = f_\mathrm{U}(T)\theta_\mathrm{U}(T) + (1 - f_\mathrm{U}(T))\theta_\mathrm{N}(T)$. In this equation,

$$f_\mathrm{U}(T) = \left(1 + \exp\left(-\frac{\Delta H_\mathrm{m}}{R}\left(\frac{1}{T} - \frac{1}{T_\mathrm{m}}\right)\right)\right)^{-1}$$

which represents the fraction of unfolded H1 globular domain; ΔH_m represents the enthalpy change of folding at the transition midpoint, and R the gas constant. $\theta_\mathrm{N}(T)$ and $\theta_\mathrm{U}(T)$ are linear baselines from the folded and unfolded states, respectively, as a function of absolute temperature, T.

Binding kinetics of H1 and ProTα. Mixing experiments were carried out with an Applied Photophysics Pi Star-180 stopped-flow spectrometer. A solution of ProTα doubly labelled with Alexa 488/594 (at positions 56 and 110) at a concentration of 2.2 nM was mixed with a solution of unlabelled H1 at variable concentrations using a 1:10 mixing ratio. The increase in acceptor fluorescence emission resulting from the compaction of ProTα on H1 binding (see Figure 2a) was used to monitor the binding reaction by exciting at 436 nm with a 10-nm bandwidth using a HgXe lamp and recording fluorescence emission using a 580-nm long-pass filter. The buffer used was TBS in the presence of 0.01% Tween 20 to minimize surface adhesion of the proteins. For each final H1 concentration between 5 nM and 100 nM, at least 80 measurements were recorded and averaged.

Simulation methods. A coarse-grained model was used for both proteins, in which each residue is represented by a single bead centred on the Cα atom. The potential energy had the functional form:

$$\begin{aligned} V = {} & \frac{1}{2}\sum_{i<N} k_\mathrm{b}(d_i - d_i^0)^2 + \frac{1}{2}\sum_{i<N-1} k_\theta(\theta_i - \theta_i^0)^2 \\ & + \sum_{i<N-2}\sum_{n=1}^{4} k_{i,n}(1 + \cos(n\phi_i - \delta_{i,n})) + \sum_{i<j}\frac{q_i q_j}{4\pi\varepsilon_\mathrm{d}\varepsilon_0}\exp\left[-\frac{d_{ij}}{\lambda_\mathrm{D}}\right] \\ & + \sum_{(i,j)\notin \mathrm{nat}} 4\varepsilon_\mathrm{pp}\left(\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right) \\ & + \sum_{(i,j)\in \mathrm{nat}} \varepsilon_{ij}\left(13\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right) \end{aligned}$$

The first two terms describe harmonic bond and angle energies, respectively, with force constants $k_\mathrm{b} = 3.16 \times 10^5$ kJ mol^{-1} nm^{-2} and $k_\theta = 6.33 \times 10^2$ kJ mol^{-1} rad^{-2}, and reference values d_i^0 and θ_i^0 taken from an extended backbone structure. The third term is a sequence-based statistical torsion potential taken from the Go model[38], and is applied to all residues, and the fourth term is a screened coulomb potential, with Debye screening length λ_D that is applied to all residues with non-zero charges q_i. ε_0 is the permittivity of free space and ε_d the dielectric constant, set here to 80. The fifth term is a generic short-range attractive potential applied to all residue pairs not identified as being part of the natively folded globular domain of H1. This interaction is characterized by a contact distance $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, in which σ_{ij} represents the residue diameters (all ~6 Å) determined from residue volumes[37], and by a contact energy ε_pp, which is the same for all such non-native residue pairs. The final term is an attractive potential applied only to the residues identified as native in the folded histone domain. The Go model[38] gives the residues that are considered native as well as the values of the parameters σ_{ij} and ε_{ij} for native pairs. For the electrostatic term, the charges are +1 for lysine and arginine, −1 for glutamate and aspartate, and +0.5 for histidine (to account for its pK_a near 6). The screening length, or Debye length, λ_D is given by

$$\lambda_\mathrm{D} = \left(\frac{\varepsilon_\mathrm{d}\varepsilon_0 k_\mathrm{B} T}{2N_\mathrm{A} e^2 I}\right)^{1/2}$$

in which k_B is the Boltzmann constant, T is the temperature, N_A is the Avogadro constant, e is the elementary charge and I is the ionic strength in molar units. The variation of ionic strength only enters the model through the screening length. Although this treatment of electrostatics is very simplified, it is consistent with the coarse-grained level of the rest of the model.

There was therefore only one free parameter to be determined (ε_pp); the same value was used for all inter- and intramolecular interactions. We varied ε_pp in order to obtain an optimal agreement with all the FRET data. This optimal value was found to be 0.16 $k_\mathrm{B}T$, or ~0.40 kJ mol^{-1}. Langevin dynamics simulations were run at a temperature of 300 K, with a friction coefficient of 0.1 ps^{-1} and a time step of 10 fs for 20 μs for each run; the mass of each bead was that of the corresponding residue. Simulations of the bound complex were started either with the molecules separated, or in an initially contacting configuration. Results from either simulation were the same, neglecting the equilibration part of the simulation. We also tested the effect of variations of the model. Using a residue-independent value of 6 Å for σ_{ij} for all residue pairs did not appreciably change the results. Similarly, using a residue-specific short-range potential similar to that devised in a previous protein interaction model[67] did not improve the agreement with experiment. However, a model with randomized or uniform charges (equal to the average) for the two proteins was unable to capture the important qualitative features of the data, in particular the difference in FRET efficiencies between the N and C termini of ProTα, and H1. This result emphasizes the dominant role of electrostatics in determining the properties of the complex.

We also considered whether the results may have been influenced by the presence of the FRET chromophores and the linkers used to covalently attach them to the protein. We therefore ran an additional set of simulations, one for each labelling combination, in which we included an explicit, coarse-grained representation of the linkers. The linkers and dyes were approximated by 5 beads (for each dye plus linker), in an unbranched chain, and with similar properties to the protein (bond lengths 3.8 Å, all bond angles 110°; the dihedral angle term was omitted). One end of the chain was bonded to the bead for the labelled residue. The motivation for the choice of 5 beads and protein-like geometry was the earlier finding that the effect of linkers on unfolded proteins can be accounted for by adding an extra 9–10 residues to the true number of residues separating the labelling positions[30,65]. The short-range interaction of the dyes with themselves and the protein was given by a Lennard–Jones term similar to that used for the other non-native interactions in the model, but the parameters were set so as to give only a short-range repulsion, with $\varepsilon = 0.001$ kJ mol^{-1}, and $\sigma = 6$ Å. Each chromophore carries a net charge of -2, which was included by adding a charge of -1 to each of the two beads furthest from the attachment point to the protein. Explicit simulations were run for each labelling combination considered in the paper.

Dissociation constants were estimated by umbrella sampling using the centre-of-mass distance between the proteins as coordinate, with harmonic umbrellas spaced between 0 and 25 nm and a force constant of 10 kJ mol^{-1} nm^{-2}. The potential of mean force $F_{\rm WHAM}(r)$ along the distance r between the centres of mass of the proteins was reconstructed using weighted histogram analysis (WHAM)[68], and the effective pair potential $F_{\rm eff}(r)$ (Extended Data Fig. 6b) was obtained from $F_{\rm eff}(r) = F_{\rm WHAM}(r) + 2k_{\rm B}T\log(r)$, in which k_B is the Boltzmann constant and T the temperature. $F_{\rm eff}$ was shifted by a constant energy so that the interaction energy at large separations was zero. The dissociation constant $K_{\rm d}$ was calculated from

$$K_{\rm d}^{-1} = 4\pi N_{\rm A} \int_0^{r_{\rm b}} \exp[-\beta F_{\rm eff}(r)] r^2 dr$$

in which $r_{\rm b}$ is the radius defining the maximum extent of the bound state (in which $F_{\rm eff}(r)$ becomes non-zero), and $\beta = 1/k_{\rm B}T$.

Conformations were initially analysed using a previously devised clustering algorithm[69], which was applied to the Hamming distances between the binary contact maps of different conformations (using a distance cut-off of 8 Å to define a contact). This algorithm identifies cluster centres as structures, i, with a high density of neighbours, ρ_i, (many structures at a short distance), but which have a large distance to the nearest structure with higher neighbour density, δ_i. The 'decision graph' consists of plotting δ_i versus ρ_i for all structures. Cluster centres should appear as points at the top right of the graph. The decision graph in this case (Extended Data Fig. 6a) shows only a single cluster. Other clustering algorithms also provided little evidence for distinct clusters, suggesting that all structures fall into a single, very broad state. We therefore used principal component analysis as a way of projecting out the structural variations. We used a set of coarse-grained inter-residue distances as the space in which to perform principal component analysis, in which only every fifth residue in the sequence was considered, and all pair distances between such residues were computed. We obtained the principal components by diagonalization of the variance–covariance matrix of this set of distances. The first three components are represented as matrices in Extended Data Fig. 6c.

Statistics and sample sizes of single molecule experiments and simulations. Minimum and average numbers of single molecules for which fluorescence has been recorded and used to build transfer efficiency histograms are indicated below every figure.

In Fig. 1c, 2 independent measurements: curves shown are the average of 60 spectra each.

In Fig. 2a, 5 independent titrations, 19 different protein concentrations, minimum number of molecules > 1,000 each (~4,000 molecules on average). Each transfer efficiency histogram constitutes an independent measurement of the affinity, because the relative populations can be determined directly from the peak integrals.

In Fig. 2b, for full-length H1: 5 independent titrations, 19 different protein concentrations, minimum number of molecules >1,000 each (~4,000 molecules on average); for H1 C-terminal disordered region: 1 titration, 19 different protein concentrations, minimum number of molecules > 2,500 each (~4,600 molecules on average); for H1 N-terminal disordered region: 1 titration, 10 different protein concentrations, minimum number of molecules > 1,100 each (~4,200 molecules on average); for H1 globular domain: 1 titration, 12 different protein concentrations, minimum number of molecules > 8,200, (~1.6×10^4 molecules on average).

In Fig. 2c, for 'Ionic strength (IS) = 180 mM': 1 titration with 8 different protein concentrations, minimum number of molecules > 2,000 each (~4,000 molecules on average); for 'IS = 205 mM': 2 titrations with an average of 14 different protein concentrations, minimum number of molecules > 2,900 each (~5,300 molecules on average); for 'IS = 240 mM': 2 titrations with an average of 10 different protein concentrations, minimum number of molecules > 3,000 each (~4,800 molecules on average); for 'IS = 290 mM': 3 titrations with an average of 12 different protein concentrations, minimum number of molecules > 2,700 each (~6,000 molecules on average); for 'IS = 330 mM': 2 titrations with an average of 8 different protein concentrations, minimum number of molecules > 3,100 each (~5,700 molecules on average); for 'IS = 340 mM': 2 titrations with an average of 6 different protein concentrations, minimum number of molecules > 950 each (~4,000 molecules on average).

In Fig. 3, ProTα$_{56/110}$ + H1: 3 independent measurements, sample size: ~5×10^6 molecules (Fig. 3a); ProTα + H1$_{0/194}$: 3 independent measurements, sample size: ~5×10^5 molecules (Fig. 3b); ProTα$_{110(\text{Alexa 594})}$ + H1$_{0(\text{Alexa 488})}$: 3 independent measurements, sample size: ~5×10^5 molecules (Fig. 3c); ProTα$_{2(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: 3 independent measurements, sample size: ~9×10^5 molecules (Fig. 3d). In Fig. 3i, 1 measurement each, minimum number of molecules > 2,900 each, ~6,800 molecules on average.

In Extended Data Fig. 4c, 1 measurement each; ProTα$_{2(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: ~7,300 molecules; ProTα$_{56(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: ~6,900 molecules; ProTα$_{110(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: ~8,700 molecules. In Extended Data Fig. 4d, 1 measurement, minimum number of molecules >3,600 (~4,500 molecules on average). In Extended Data Fig. 4e, 1 measurement, minimum number of molecules >800 (~1,400 molecules on average).

In Extended Data Fig. 5, 1 experiment each; ProTα$_{56/110}$: ~7,500 molecules; ProTα$_{56/110}$ + H1: ~3,000 molecules; H1$_{0/113}$: ~2,000 molecules; H1$_{0/113}$ + ProTα: ~3,300 molecules; ProTα$_{2(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: ~7,400; ProTα$_{110(\text{Alexa 594})}$ + H1$_{194(\text{Alexa 488})}$: ~8,700.

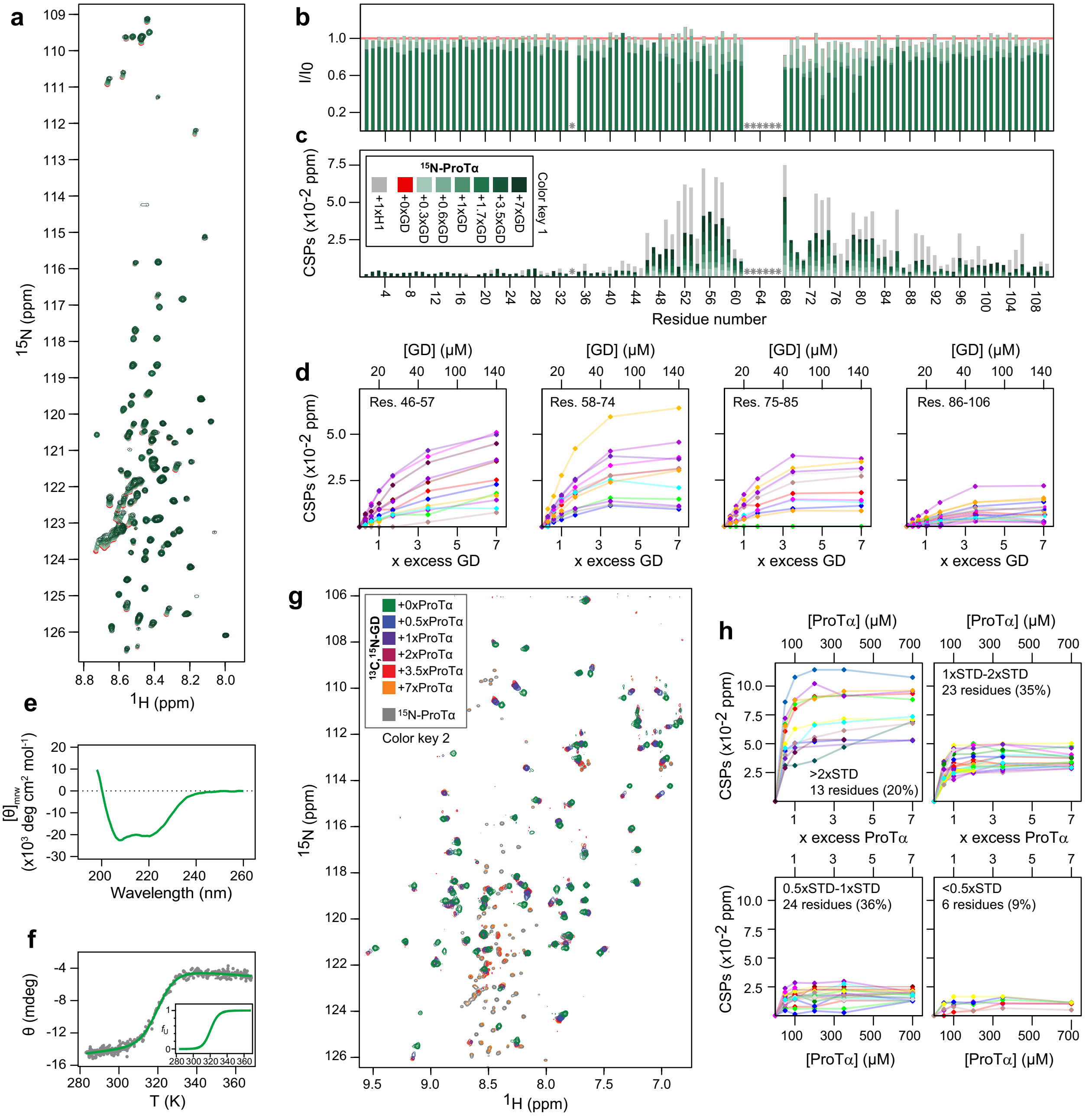
The uncertainty of the FRET efficiencies estimated from simulations was determined by block analysis in a similar fashion to a previously described method[70], in which the error around the mean is estimated from statistically independent blocks of simulation data.

Code availability. A custom module for Mathematica (Wolfram Research) used for the analysis of single-molecule fluorescence data is available upon request.

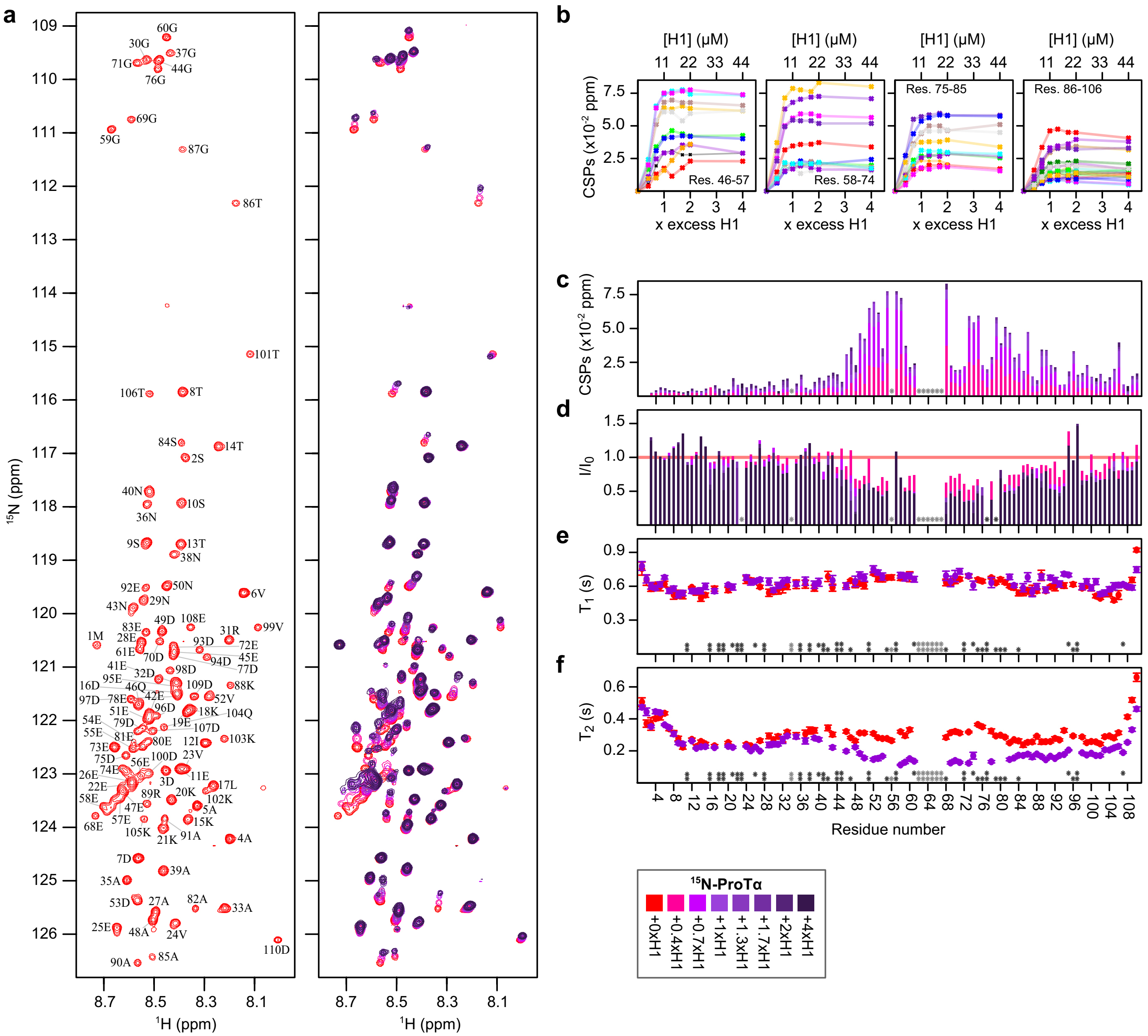
Data availability. All data supporting the findings of this study are available within the paper and its Supplementary Information. The raw data are available from the corresponding authors upon reasonable request. Source Data for Fig. 2 and Extended Data Table 2 are provided with the paper. The chemical shift assignments of ProTα alone and in complex with H1 have been deposited to the Biological Magnetic Resonance Bank (http://www.bmrb.wisc.edu/) under accession numbers 27215 and 27216, respectively.

51. Scott, K. A., Steward, A., Fowler, S. B. & Clarke, J. Titin: a multidomain protein that behaves as the sum of its parts. *J. Mol. Biol.* **315,** 819–829 (2002).
52. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
53. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59,** 687–696 (2005).
54. Orekhov, V. Y. & Jaravine, V. A. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. *Prog. Nucl. Magn. Reson. Spectrosc.* **59,** 271–292 (2011).
55. Fedyukina, D. V. *et al.* Contribution of long-range interactions to the secondary structure of an unfolded globin. *Biophys. J.* **99,** L37–L39 (2010).
56. Gibbs, S. J. & Johnson, C. S. Jr. A PFG NMR experiment for accurate diffusion and flow studies in the presence of eddy currents. *J. Magn. Reson.* **93,** 395–402 (1991).
57. Farrow, N. A. *et al.* Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33,** 5984–6003 (1994).
58. Benke, S. *et al.* The assembly dynamics of the cytolytic pore toxin ClyA. *Nat. Commun.* **6,** 6198 (2015).
59. Müller, B. K., Zaychikov, E., Bräuchle, C. & Lamb, D. C. Pulsed interleaved excitation. *Biophys. J.* **89,** 3508–3522 (2005).
60. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods* **3,** 891–893 (2006).
61. Schuler, B. Application of single molecule Förster resonance energy transfer to protein folding. *Methods Mol. Biol.* **350,** 115–138 (2007).
62. Kellner, R. *et al.* Single-molecule spectroscopy reveals chaperone-mediated expansion of substrate protein. *Proc. Natl Acad. Sci. USA* **111,** 13355–13360 (2014).
63. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437,** 55–75 (1948).
64. Gopich, I. V., Nettels, D., Schuler, B. & Szabo, A. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J. Chem. Phys.* **131,** 095102 (2009).
65. Borgia, A. *et al.* Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* **138,** 11714–11726 (2016).
66. Dertinger, T. *et al.* Two-focus fluorescence correlation spectroscopy: a new tool for accurate and absolute diffusion measurements. *ChemPhysChem* **8,** 433–443 (2007).

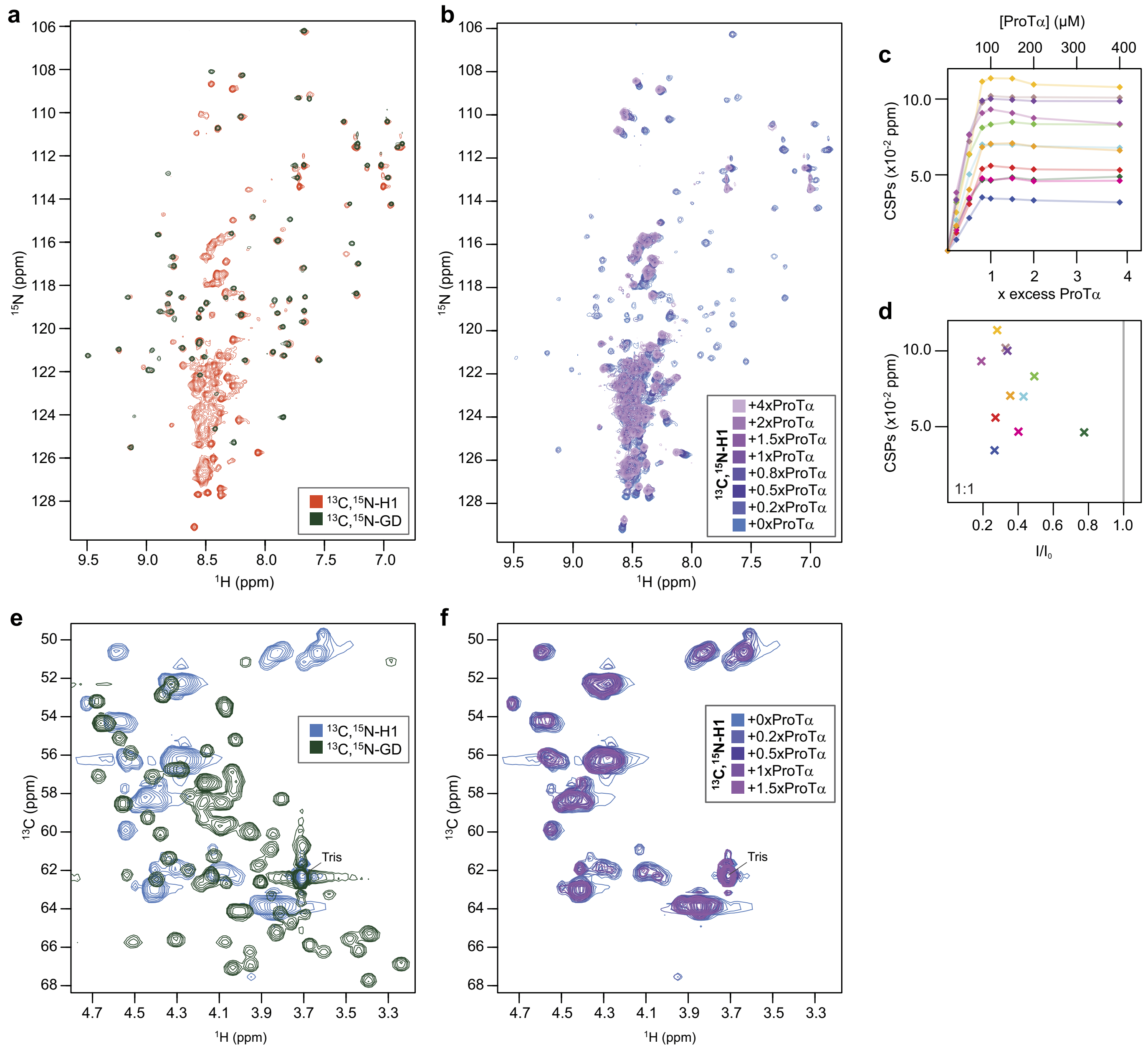
67. Kim, Y. C. & Hummer, G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* **375,** 1416–1433 (2008).
68. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13,** 1011–1021 (1992).
69. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. Science **344,** 1492–1496 (2014).
70. Flyvbjerg, H. & Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **91,** 461–466 (1989).
71. Kapanidis, A. N. *et al.* Alternating-laser excitation of single molecules. *Acc. Chem. Res.* **38,** 523–533 (2005).
72. Hoffmann, A. *et al.* Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl Acad. Sci. USA* **104,** 105–110 (2007).



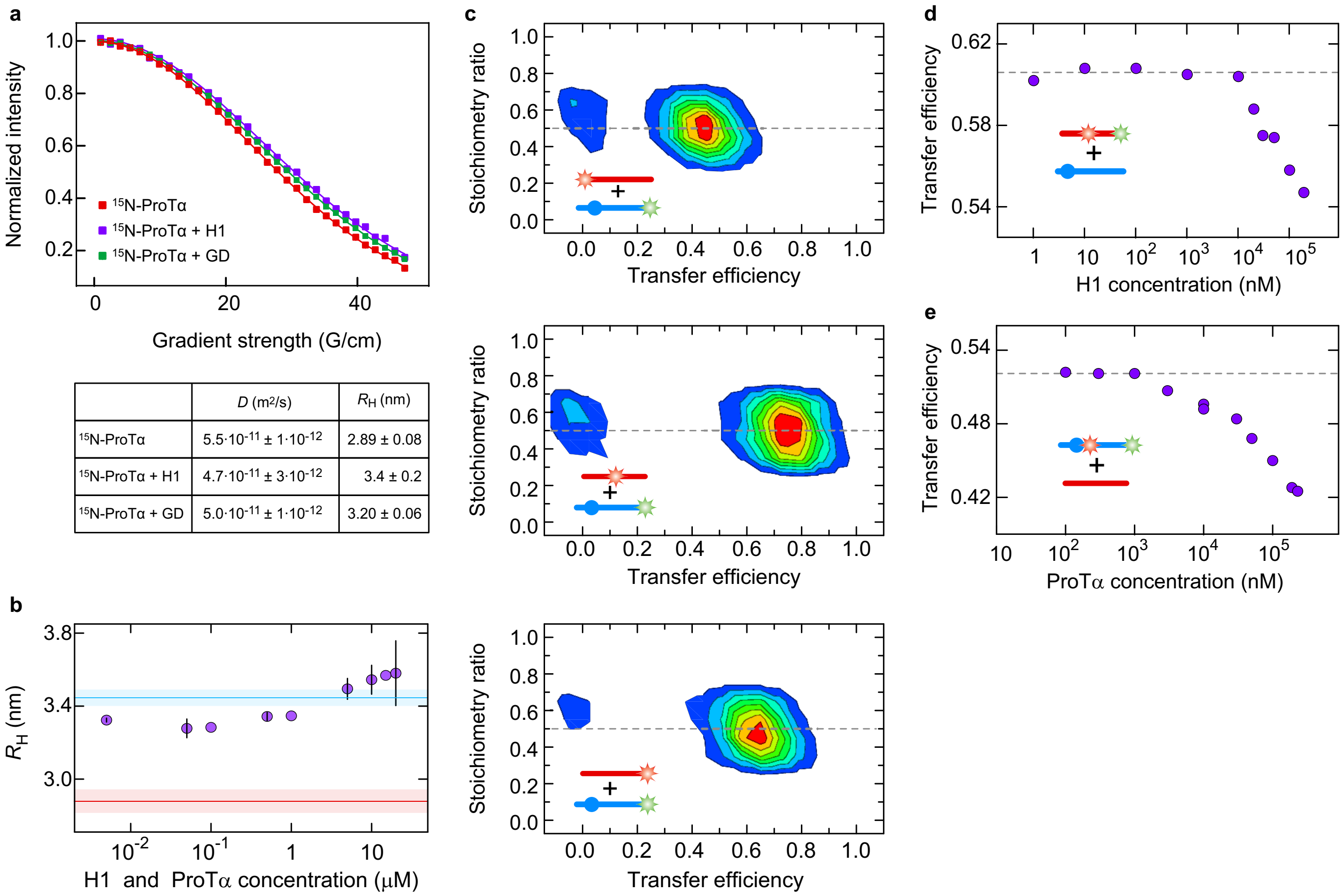
Extended Data Figure 1 | Titrations of ProTα and H1 globular domain. **a**, Titration of ^{15}N-ProTα with zero to sevenfold molar addition of the H1 globular domain followed by ^{1}H–^{15}N HSQC spectra; $n = 2$ repeats of this measurement yielded consistent results. **b**, Peak intensity ratios for assigned residues of ProTα relative to the free state induced by zero to 1.7-fold molar addition of the H1 globular domain ($n = 2$). **c**, Weighted backbone CSPs per residue of ProTα induced by zero to sevenfold molar addition of the H1 globular domain ($n = 2$). For comparison, CSPs of ProTα with equimolar addition of H1 are shown in grey ($n = 5$). In **a**–**c**, ‘colour key 1’ applies; grey stars, prolines and unassigned residues. **d**, ProTα CSPs plotted against concentration and times excess of the H1 globular domain relative to the free state for residues 46–106 upon zero to sevenfold molar addition of the H1 globular domain. Curves corresponding to individual residues are shown in different colours for clarity. **e**, Far-UV circular dichroism spectrum of the H1 globular domain. **f**, Thermal denaturation of the H1 globular domain followed by the change in ellipticity at 222 nm ($T_m = 320.5 \pm 0.3$ K, $\Delta H_m = -44 \pm 2$ kcal mol^{-1}). Inset in **f** shows fraction of unfolded H1 globular domain (f_U) as a function of temperature. **g**, Titration of 100 μM ^{13}C–^{15}N- H1 globular domain with zero to sevenfold molar addition of ProTα followed by ^{1}H–^{15}N HSQC spectra. Peak intensities gradually decrease during the titration. At 3.5- and 7-fold molar excess ProTα, natural abundance peaks of free ProTα appear. ^{1}H–^{15}N HSQC spectrum of ^{15}N-ProTα is shown in grey for comparison. **h**, Weighted backbone CSPs of the H1 globular domain plotted against concentration and times excess of ProTα relative to the free state on zero to sevenfold molar addition of ProTα. A total of 66 (unassigned) amide backbone peaks were followed and grouped according to the standard deviation (STD) of the CSPs (1 s.d. = 0.0254 p.p.m.). Of these, 55% had CSPs larger than 1 STD.



Extended Data Figure 2 | Titration of ^{15}N-ProTα with H1. a, ^{1}H–^{15}N HSQC spectrum of 11 μM free ^{15}N-ProTα with assigned residues labelled (left) and titrated with zero to fourfold molar addition of H1 (right); $n = 5$ individual repeats of this measurement yielded consistent results. **b**, Weighted backbone CSPs of ProTα (residues 46–106) relative to the free state on zero to fourfold molar addition of H1, plotted against concentration and times excess of H1. Curves corresponding to individual residues are shown in different colours for clarity. **c**, **d**, CSPs (**c**) and peak intensity ratios (**d**) for assigned residues of ProTα induced by zero to fourfold molar addition of H1 (bar colours correspond to key); $n = 5$ for both. **e**, T_1 ^{15}N relaxation times of free (red) and H1-bound (purple) ^{15}N-ProTα. $\langle T_1 \rangle = 610$ ms (free) and 636 ms (complex); $n = 2$ individual repeats of this measurement yielded consistent results. **f**, T_2 ^{15}N relaxation times of free (red) and H1-bound (purple) ^{15}N-ProTα. $\langle T_2 \rangle = 302$ ms (free) and 217 ms (complex). In **c**–**f**, light grey stars, prolines and unassigned residues; dark grey stars, overlap and/or insufficient data quality. Circles in **e** and **f** are mean values from $n = 3$ consecutive data acquisitions on the same samples, errors are s.d.

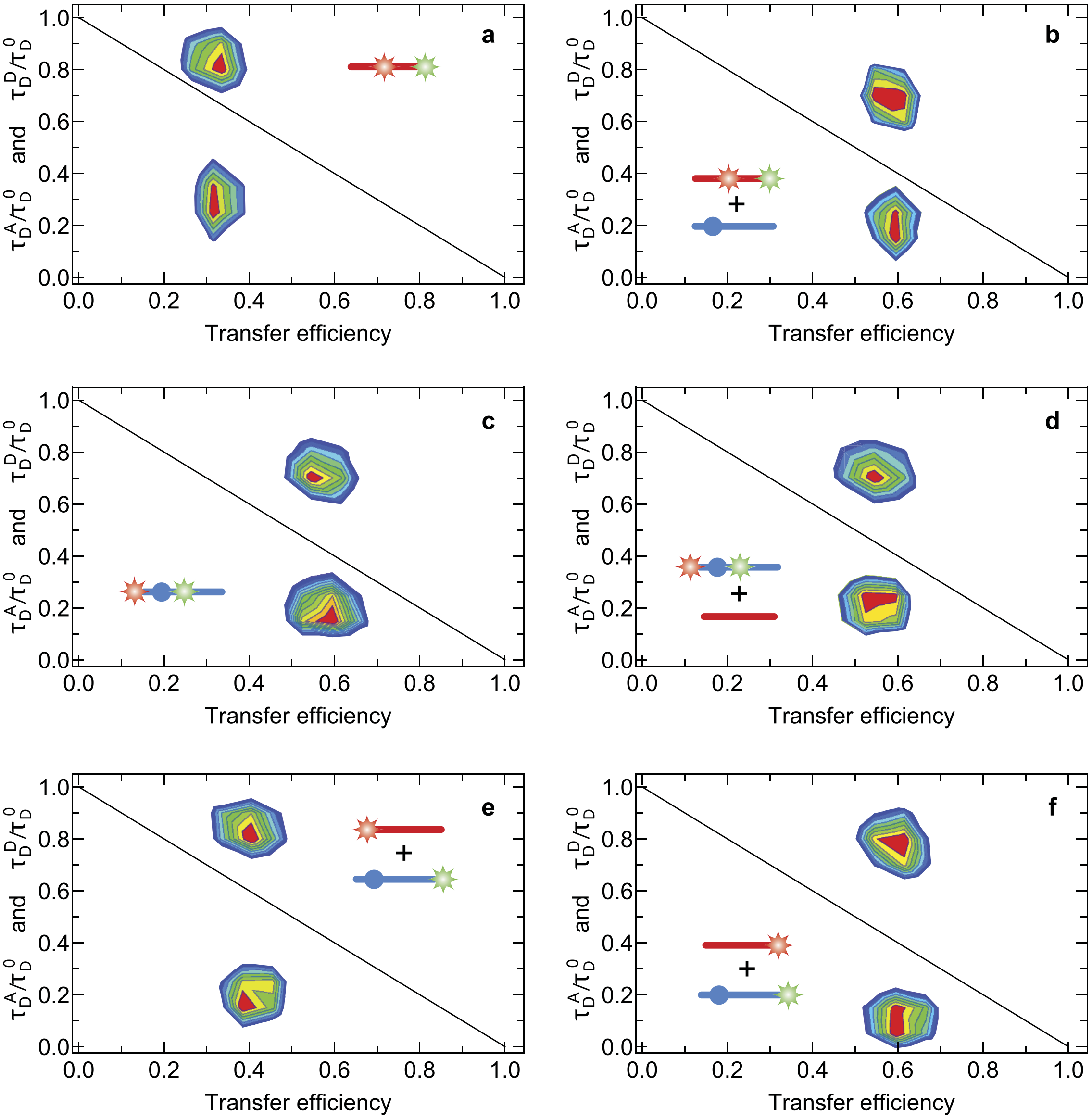


Extended Data Figure 3 | Titration of ^{13}C–^{15}N-H1 with ProTα. **a**, ^{1}H–^{15}N HSQC spectra of the free ^{13}C–^{15}N-H1 globular domain (dark green) and free ^{13}C–^{15}N-H1 (orange). The majority of the amide peaks of the H1 globular domain overlap with the more dispersed peaks from H1, indicating the similarity in structure of the H1 globular domain in isolation and within H1. **b**, Titration followed by ^{1}H–^{15}N HSQC spectra of ^{13}C–^{15}N-H1 with zero to fourfold molar addition of ProTα. Data acquired on His$_6$-tagged H1; $n = 2$ individual repeats of this measurement yielded consistent results. **c**, CSPs relative to free H1 of 11 traceable H1 amide backbone peaks from the intrinsically disordered region (based on overlay with ^{1}H–^{15}N HSQC spectra of the H1 globular domain (in **a**)) on zero to fourfold molar addition of ProTα plotted against concentration and times excess. Curves corresponding to individual residues are shown in different colours for clarity. **d**, CSPs plotted against peak intensity ratios relative to the free state of H1 of the 11 H1 amides at 1× excess of ProTα. Colours correspond to those in **c**. **e**, Overlay of the Cα–Hα region from ^{1}H–^{13}C HSQC spectra of free ^{13}C–^{15}N-H1 (blue) and the ^{13}C–^{15}N-H1 globular domain (green). The H1 ^{1}H–^{13}C HSQC spectrum is dominated by intense clusters of peaks not present in the H1 globular domain spectrum, consistent with the large fraction of residue repeats in the H1 disordered regions. **f**, Cα–Hα region of ^{13}C–^{15}N-H1 on titration with ProTα. The lack of detectable changes in Cα–Hα resonances is consistent with the absence of secondary structure induction in the disordered regions of H1 on binding.

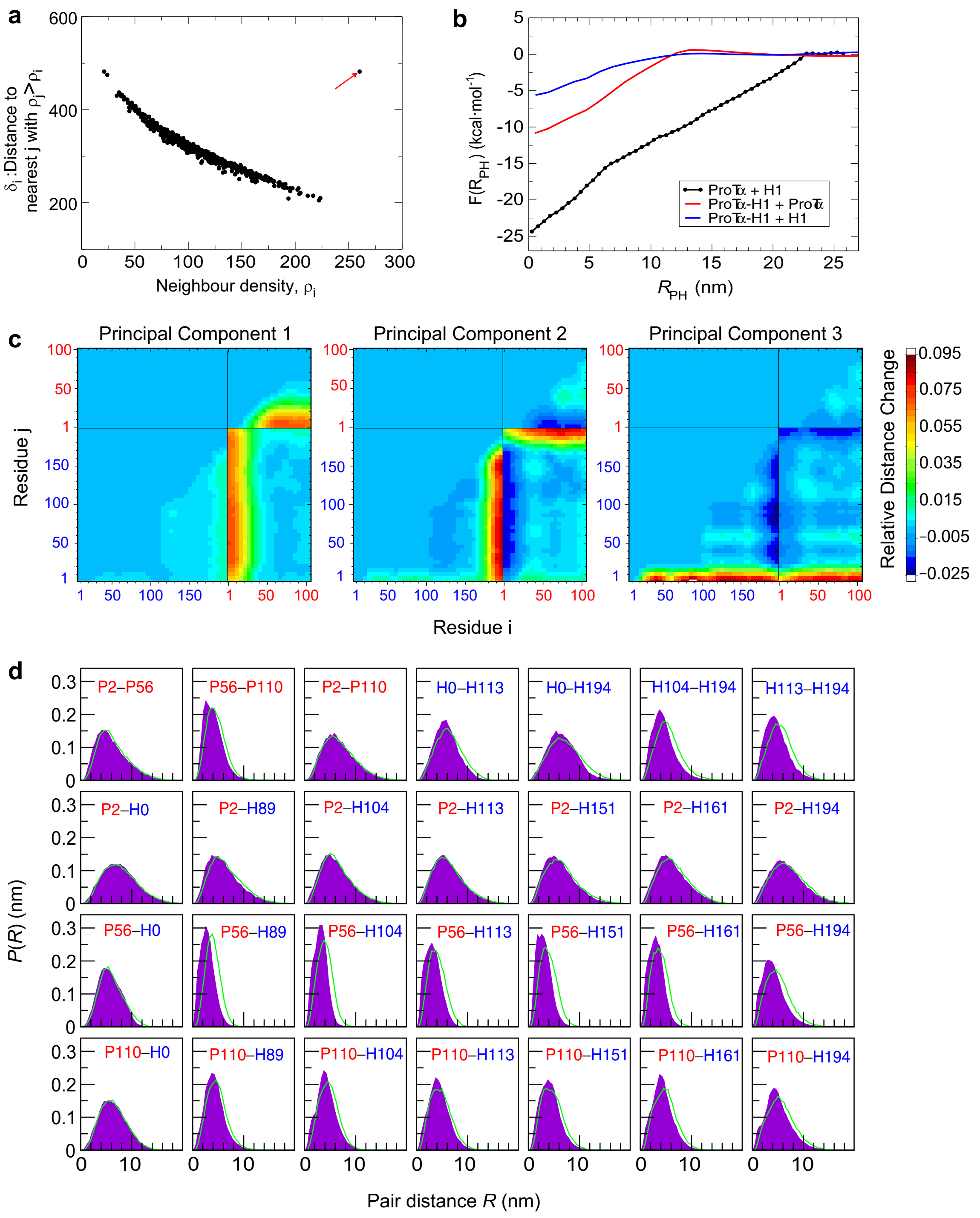


	D (m²/s)	R_H (nm)
^{15}N-ProTα	$5.5 \cdot 10^{-11} \pm 1 \cdot 10^{-12}$	2.89 ± 0.08
^{15}N-ProTα + H1	$4.7 \cdot 10^{-11} \pm 3 \cdot 10^{-12}$	3.4 ± 0.2
^{15}N-ProTα + GD	$5.0 \cdot 10^{-11} \pm 1 \cdot 10^{-12}$	3.20 ± 0.06

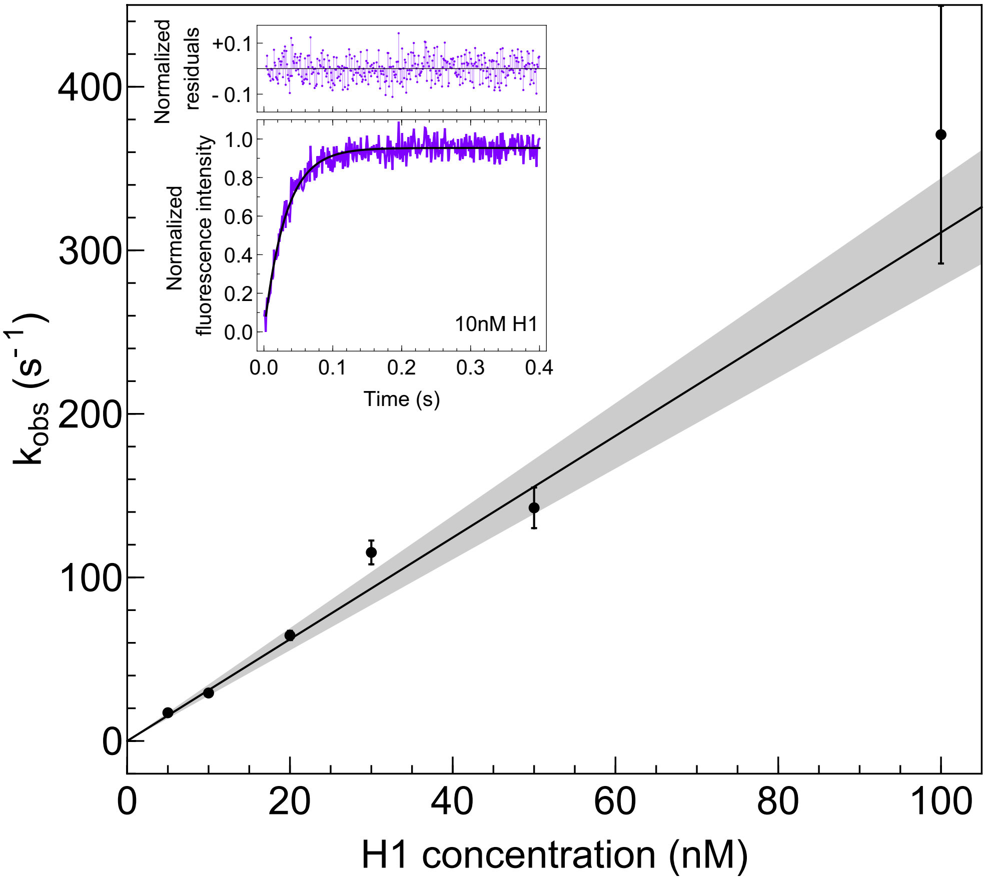
Extended Data Figure 4 | Hydrodynamic radii and stoichiometry of the H1–ProTα complex. a, R_H of free and bound ^{15}N-ProTα (100 μM) determined with pulsed-field gradient NMR at 283 K. The signal decays of free ^{15}N-ProTα (red), with H1 at a 1:1 molar ratio (purple) and with the H1 globular domain at a 1:7 molar ratio (green) as a function of gradient strength, together with corresponding fits and a table of the diffusion coefficients and resulting R_H values. **b**, R_H measured by two-focus fluorescence correlation spectroscopy at 295 K. Lines show the mean R_H from $n = 2$ independent measurements of $H1_0$ (blue) and ProTα_2 (red) labelled with Alexa 594 in the absence of the binding partner. Symbols represent the mean R_H from $n = 2$ independent measurements of labelled ProTα (5 nM) in the presence of equimolar concentrations of unlabelled ProTα and unlabelled H1. Error bars or shaded bands, s.d. **c**, Stoichiometry ratio[71] versus transfer efficiency plots from intermolecular single-molecule FRET measurements of ProTα_2 + $H1_{194}$ (top), ProTα_{56} + $H1_{194}$ (middle), and ProTα_{110} + $H1_{194}$ (bottom); pictograms in panels indicate labelling locations. A stoichiometry ratio of 0.5 indicates a 1:1 complex. The peaks at $E \approx 0$ originate from molecules or complexes that lack an acceptor dye and remain after filtering for donor-only fluorescence bursts based on pulsed-interleaved excitation. **d**, **e**, Transfer efficiency changes at a large excess of unlabelled binding partner for FRET-labelled ProT$\alpha_{56/110}$ (**d**) and $H1_{104/194}$ (**e**). See Methods for further information on statistics.



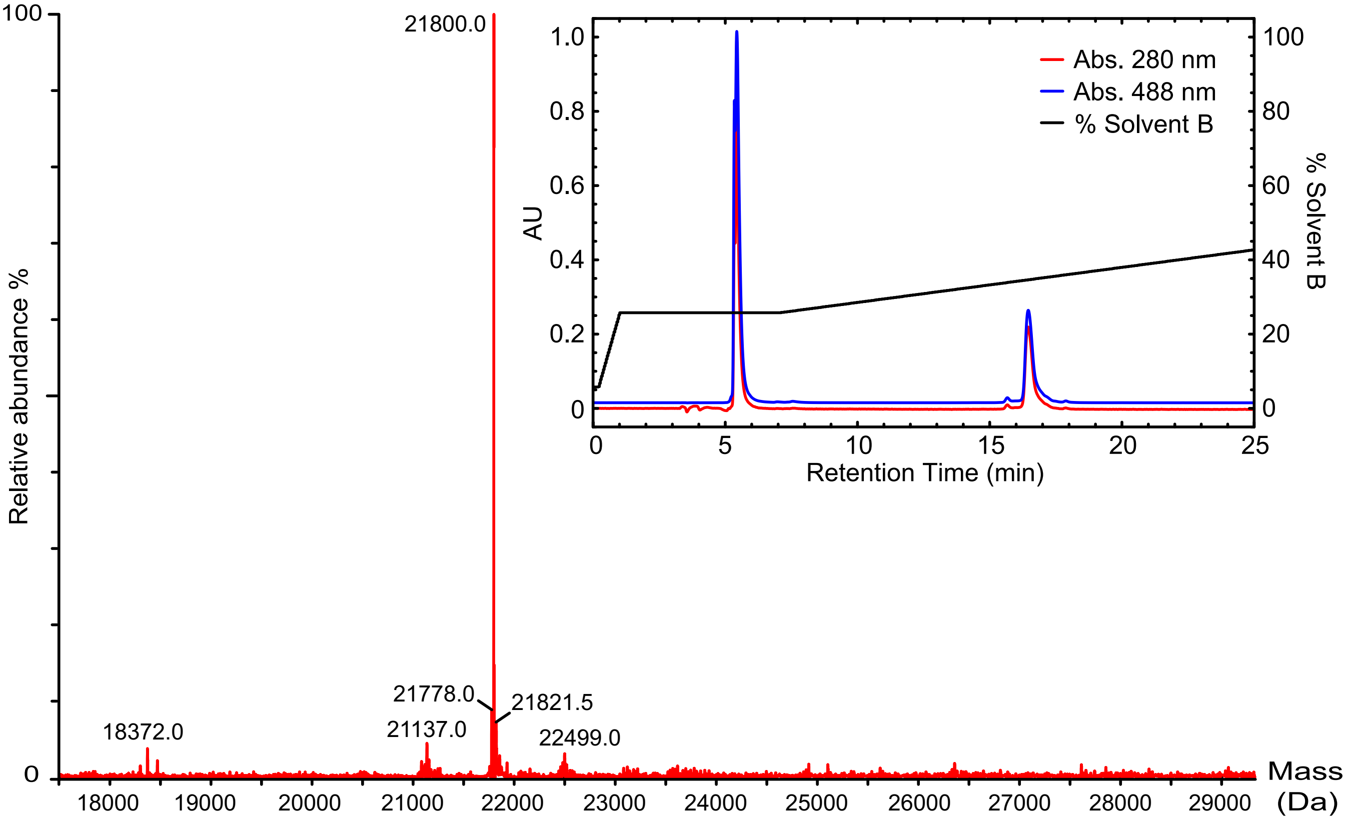
Extended Data Figure 5 | Fluorescence lifetime analysis. a–f, Plots of the fluorescence lifetimes of Alexa 488 donor (τ_D^D) and Alexa 594 acceptor (τ_D^A) normalized by the intrinsic donor lifetime (τ_D^0) versus the ratiometric transfer efficiency *E* (calculated from the number of donor and acceptor photon counts), as a diagnostic for the presence of a broad distance distribution rapidly sampled during the time of a fluorescence burst[28,33,34]. If fluctuations in transfer efficiency occur on a timescale between the donor fluorescence lifetime (~4 ns) and the burst duration (~1 ms), the normalized donor lifetimes cluster above—and the acceptor lifetimes below—the solid diagonal line expected for a single fixed distance, as previously observed for intrinsically disordered proteins[34,72]. The large deviation from the diagonal observed for both unbound and bound ProTα and H1 supports the presence of broad and rapidly sampled distance distributions. **a**, ProT$\alpha_{56/110}$; **b**, ProT$\alpha_{56/110}$ + unlabelled H1; **c**, H$1_{0/113}$; **d**, H$1_{0/113}$ + unlabelled ProTα; **e**, ProTα_2 + H1_{194}; and **f**, ProTα_{110} + H1_{194}. All variants labelled with Alexa 488 (green) and/or Alexa 594 (red) as indicated by the pictograms in the figure panels.



Extended Data Figure 6 | Simulation results. a, Decision graph using the Rodriguez–Laio clustering algorithm[69], showing only a single density maximum distant from other density maxima (that is, a single distinct cluster). **b**, Free energy of association from simulation for ProTα and H1 along the distance between their centres of mass, R_{PH}, yielding a K_{d} of 7 fM (black curve). Blue and red curves are the free energies for addition of a second H1 or a second ProTα, respectively, to an existing H1–ProTα complex. **c**, Principal component vectors shown as contact maps. Colours indicate the increase or decrease in each pair distance for that principal component, relative to the other distances. ProTα and H1 residue numbers are indicated in red and blue, respectively. Each principal component describes a feature of the chain arrangement: principal component 1, for example, captures the presence or absence of interactions between the ProTα N terminus and H1. **d**, Intramolecular (top row) and intermolecular (bottom three rows) distributions of distances corresponding to FRET labelling sites, within the H1–ProTα complex. P and H numbers refer to ProTα and H1 residues, respectively. Filled distributions, simulations without explicit chromophores; green lines, simulations with explicit chromophores.



Extended Data Figure 7 | Kinetics of H1–ProTα binding measured by stopped flow. FRET-labelled ProT$\alpha_{56/110}$ was mixed rapidly with unlabelled H1 in TBS buffer, and the resulting increase in acceptor fluorescence was monitored. Inset, example at 10 nM H1 with single-exponential fit and residuals shown above (see Methods for details). Decay rates were obtained from single-exponential fits, with an instrument dead time of 3 ms. Standard errors for each H1 concentration were obtained using bootstrapping. The observed rate, k_{obs}, is shown as a function of H1 concentration (c_{H1}); for H1 concentrations between 10 and 100 nM—for which pseudo-first order conditions apply (ProTα concentration after mixing was 2 nM)—the observed rates were fit with $k_{obs} = k_{on}c_{H1} + k_{off} = k_{on}c_{H1} + k_{on}K_d$, using the independently determined K_d of 2.1 pM (Extended Data Table 2). The fit yields a bimolecular association rate coefficient of $k_{on} = 3.1 \pm 0.1 \times 10^9\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$ and an apparent dissociation rate coefficient of $k_{off} = 6.5 \pm 3.1 \times 10^{-3}\ \mathrm{s}^{-1}$. The grey area represents the 95% confidence band.



Extended Data Figure 8 | Example of the quality of the H1 preparation. Electrospray ionization mass spectrum of H1(T161C) labelled with Alexa 488 (calculated mass 21,800 Da) and preparative reversed-phase HPLC (Vydac C4) chromatogram (inset) showing absorption at 280 nm (red) and 488 nm (blue) and the elution gradient from solvent A (H_2O + 0.1% TFA) to solvent B (100% acetonitrile) (black), illustrating the high purity of the sample. The peak at approximately 5.5 min corresponds to free Alexa 488, and the peak at approximately 16.8 min to H1(T161C) labelled with Alexa 488.

Extended Data Table 1 | Sequences of protein constructs and fluorescently labelled variants of H1 and ProTα

H1 (+53)	0 23 **C**TENSTSAPAAKPKRAKASKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQI 89 96 104 113 KLSIKRLVTTGVLKQTKGVGA**S**GSFRLAKSDEPKKS**V**AFKKTKKE**I**KKVATPKKASKPKKAASKAPTK 151 161 194 KPKATPVKKAKKKLA**A**TPKKAKKPK**T**VKAKPVKASKPKKAKPVKPKAKSSAKRAGKKK**G**GPR
H1 †CTR (+39)	103 SVAFKKTKKEIKKVATPKKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASK 193 PKKAKPVKPKAKSSAKRAGKKKGGPR
H1 ‡NTR (+18)	0 G**C**TENSTSAPAAKPKRAKASKKSTDHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQI 113 KLSIKRLVTTGVLKQTKGVGASGSFRLAKSDEPKKSVAFKKTKKE**I**
GD (+9)	23 DHPKYSDMIVAAIQAEKNRAGSSRQSIQKYIKSHYKVGENADSQIKLSIKRLVTTGVLKQTKGVGASGS 96 FRLAK
ProTα (-44)	2 56 GPS**D**AAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEE**E**EEGGEEE 110 EEEEEGDGEEEDGDEDEEAESATGKRAAEDDEDDDVDTKKQKTDED**D**

H1				**ProTα**				
Singly labeled		**Doubly labeled**		**Singly labeled**		**Doubly labeled**		
Alexa 488	Alexa 594	Alexa 488/ Alexa 594 (R_0=5.4 nm)§	Cy3B/ Abberior* 635 (R_0 = 5.9 nm)§	Alexa 488	Alexa 594	Alexa 488/ Alexa 594 (R_0 = 5.4 nm)§	Cy3B/ Abberior* 635 (R_0 = 5.9 nm)§	Atto 550/ Atto 647N (R_0 = 6.6 nm)§
C0, S89C, V104C, I113C, A151C, T161C, G194C	C0, S89C, G194C	C0/I113C, C0/G194C, V104C/G194C, I113C/G194C	V104C/G194C	D110C	D2C, E56C, D110C	D2C/E56C, E56C/D110C, D2C/D110C	E56C/D110C	E56C/D110C

Top, sequences of wild-type H1, H1 fragments used, and wild-type ProTα. Residues in bold yellow are positions mutated to Cys for fluorophore conjugation. Residues in red are remain after proteolytic cleavage of the HisTag with thrombin (Gly-Gly-Pro-Arg or Gly-Cys) or HRV-3C (Gly-Pro). Note that the wild-type sequence of H1 starts with Thr1 and ends with Lys193; the preceding Cys residue (0) was added for labelling. The underlined part of the H1 sequence indicates the globular domain (GD), identified based on a sequence alignment with the *Gallus gallus* homologue[20] (RCSB Protein Data Bank access code: 1HST, 82% sequence identity). Surface-exposed residues in GD (as shown in Fig. 1a and 4b) are shaded in light blue. The net charge of each variant is indicated in parentheses Bottom, labelled variants of H1 and ProTα.

†CTR, C-terminal disordered region.

‡NTR, N-terminal disordered region including H1 globular domain.

§Förster radius of the corresponding dye pair.

Extended Data Table 2 | Binding affinities, molecular dimensions and reconfiguration times of fluorescently labelled H1 and ProTα

Affinity of ProTα 56/110 Alexa 488/594 for full-length H1 and fragments in TBS				Affinities in TBS 205 mM		Intramolecular transfer efficiencies and distances in TBS 205 mM				
Ionic strength (mM)	**K_d (nM)**	**H1 fragm.**	**K_d (TBS 165 mM)**	**ProTα**	**‖K_d (nM)**	**ProTα**	***E* unbound**	***E* bound**	**$R_{unbound}$ (nm)**	**R_{bound} (nm)**
†165	$(2.1^{+1.1}_{-0.8})\cdot 10^{-3}$	‡CTR	40^{+6}_{-4} pM	D2C/D110C Alexa 488/594	2.0 ± 0.13	E56C/D110C Alexa 488/594	0.36	0.54	$7.5^{+1.0}_{-0.3}$	$5.8^{+0.7}_{-0.1}$
180	$(37 \pm 5)\cdot 10^{-3}$					E56C/D110C Cy3B/Abb.*635	0.41	0.56	$7.6^{+0.6}_{-0.5}$	6.2± 0.4
205	1.0 ± 0.1	‡NTR	173^{+29}_{-28} nM	E56C/D110C Cy3B/Abb.*635	1.0 ± 0.10	E56C/D110C Atto 550/647N	0.45	0.59	$8.1^{+0.6}_{-0.5}$	$6.7^{+0.5}_{-0.4}$
240	25 ± 3			E56C/D110C Atto 550/647N	3.1 ± 0.20	D2C/D110C Alexa 488/594	0.18	0.33	$10.6^{+1.6}_{-1.1}$	$7.9^{+0.7}_{-0.6}$
290	$(2.3 \pm 1.5)\cdot 10^{2}$	‡GD	$1.9^{+0.3}_{-0.3}$ μM§	**H1**		**H1**	**In TBS 165 mM**			
						V104C/G194C Alexa 488/594	0.15	0.53	$11.4^{+2.1}_{-1.4}$	5.9± 0.4
330	$(1.4 \pm 0.4)\cdot 10^{3}$			V104C/G194C Alexa 488/594	3.5 ± 0.23	V104C/G194C Cy3B/Abb.*635	0.18	0.52	$11.6^{+1.8}_{-1.2}$	$6.5^{+0.5}_{-0.4}$
340	$(4.0 \pm 1.8)\cdot 10^{3}$					I113C/G194C Alexa 488/594	0.23	0.58	$9.5^{+1.1}_{-0.9}$	5.6± 0.4

ProTα A-594 **H1 A-488**	**τ_r (ns)**			**Labeled Protein (Alexa 488/ Alexa 594)**	**τ_r (ns)**	
	D2C	**E56C**	**D110C**		**unbound**	**bound**
H1 C0	180^{+19}_{-16}	191^{+22}_{-19}	169^{+19}_{-16}	ProTα E56C/D110C	29±2	102^{+3}_{-2}
H1 I113C	121^{+13}_{-11}					
H1 A151C	124^{+13}_{-12}	98^{+16}_{-2}		ProTα D2C/E56C	33±2	66±2
H1 G194C	156^{+16}_{-14}		142^{+19}_{-15}			
A-488 **A-594**				ProTα D2C/D110C	78^{+15}_{-9}	133^{+10}_{-7}
H1 G194C			120^{+13}_{-12}	H1 I113C/G194C	118^{+24}_{-14}	143^{+5}_{-4}

Top left, affinities of labelled ProTα for unlabelled H1 at different ionic strengths (IS), and for H1 fragments at 165 mM IS. Uncertainties for the IS dependence are standard errors estimated from independent titrations (see Statistics in Methods). Top centre, binding affinities of ProTα and H1 labelled with different dye pairs for the respective unlabelled partner. Top right, transfer efficiencies and average distances of ProTα and H1 labelled with different dye pairs in the unbound ($R_{unbound}$) and bound state (R_{bound}) (with the respective unlabelled partner; Abb.*635 = Abberior Star 635). Uncertainties in distance are based on an estimated systematic error of ± 0.05 in the transfer efficiency from instrument calibration for the different dye pairs. Bottom left, intermolecular reconfiguration times for the complex of donor-labelled H1 (Alexa 488) and acceptor-labelled ProTα (Alexa 594) and vice versa. Bottom right, reconfiguration times of doubly labelled ProTα and H1 (unbound and bound). Uncertainties estimated by propagating the systematic error on the transfer efficiency (± 0.05).

†Uncertainty at 165 mM (see Methods for details).

‡For H1 fragment sequences, see Table 1).

§Apparent K_d from fraction of all bound species.

||Uncertainties based on dilution errors.

LETTER

doi:10.1038/nature25792

An absorption profile centred at 78 megahertz in the sky-averaged spectrum

Judd D. Bowman[1], Alan E. E. Rogers[2], Raul A. Monsalve[1,3,4], Thomas J. Mozdzen[1] & Nivedita Mahesh[1]

After stars formed in the early Universe, their ultraviolet light is expected, eventually, to have penetrated the primordial hydrogen gas and altered the excitation state of its 21-centimetre hyperfine line. This alteration would cause the gas to absorb photons from the cosmic microwave background, producing a spectral distortion that should be observable today at radio frequencies of less than 200 megahertz[1]. Here we report the detection of a flattened absorption profile in the sky-averaged radio spectrum, which is centred at a frequency of 78 megahertz and has a best-fitting full-width at half-maximum of 19 megahertz and an amplitude of 0.5 kelvin. The profile is largely consistent with expectations for the 21-centimetre signal induced by early stars; however, the best-fitting amplitude of the profile is more than a factor of two greater than the largest predictions[2]. This discrepancy suggests that either the primordial gas was much colder than expected or the background radiation temperature was hotter than expected. Astrophysical phenomena (such as radiation from stars and stellar remnants) are unlikely to account for this discrepancy; of the proposed extensions to the standard model of cosmology and particle physics, only cooling of the gas as a result of interactions between dark matter and baryons seems to explain the observed amplitude[3]. The low-frequency edge of the observed profile indicates that stars existed and had produced a background of Lyman-α photons by 180 million years after the Big Bang. The high-frequency edge indicates that the gas was heated to above the radiation temperature less than 100 million years later.

Observations with the Experiment to Detect the Global Epoch of Reionization Signature (EDGES) low-band instruments, which began in August 2015, were used to detect the absorption profile. Each of the two low-band instruments consists of a radio receiver and a zenith-pointing, single-polarization dipole antenna. Spectra of the brightness temperature of the radio-frequency sky noise, spatially averaged over the large beams of the instruments, were recorded between 50 MHz and 100 MHz. Raw spectra were calibrated, filtered and integrated over hundreds of hours. Automated measurements of the reflection coefficients of the antennas were performed in the field. Other measurements were performed in the laboratory, including of the noise waves and reflection coefficients of the low-noise amplifiers and additional calibration constants. Details of the instruments, calibration, verification and model fitting are described in Methods.

In Fig. 1 we summarize the detection. It shows the spectrum observed by one of the instruments and the results of model fits. Galactic synchrotron emission dominates the observed sky noise, yielding a power-law spectral profile that decreases from about 5,000 K at 50 MHz to about 1,000 K at 100 MHz for the high Galactic latitudes shown. Fitting and removing the Galactic emission and ionospheric contributions from the spectrum using a five-term, physically motivated foreground model (equation (1) in Methods) results in a residual with a root-mean-square (r.m.s.) of 0.087 K. The absorption profile is found by fitting the integrated spectrum with the foreground model and a model for the 21-cm signal simultaneously. The best-fitting 21-cm model yields a symmetric U-shaped absorption profile that is centred at a frequency of 78 ± 1 MHz and has a full-width at half-maximum of 19^{+4}_{-2} MHz, an amplitude of $0.5^{+0.5}_{-0.2}$ K and a flattening factor of $\tau = 7^{+5}_{-3}$ (where the bounds provide 99% confidence intervals including estimates of systematic uncertainties; see Methods for model definition). Uncertainties in the parameters of the fitted profile are estimated from statistical uncertainty in the model fits and from systematic differences between the various validation trials that were performed using observations from both instruments and several different data cuts. The 99% confidence intervals that we report are calculated as the outer bounds of (1) the marginalized statistical 99% confidence intervals from fits to the primary dataset and (2) the range of best-fitting values for each parameter across the validation trials. Fitting with both the foreground and 21-cm models lowers the residuals to an r.m.s. of 0.025 K. The fit shown in Fig. 1 has a signal-to-noise ratio of 37, calculated as the best-fitting amplitude of the profile divided by the statistical uncertainty of the amplitude fit, including the covariance between model parameters. Additional analyses of the

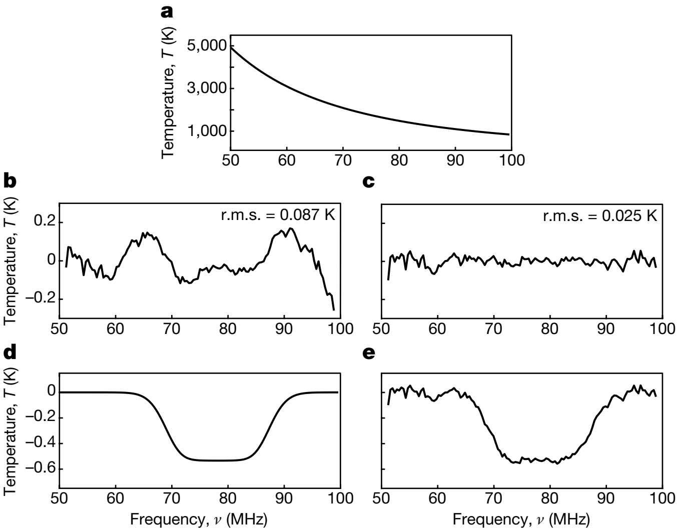


Figure 1 | Summary of detection. a, Measured spectrum for the reference dataset after filtering for data quality and radio-frequency interference. The spectrum is dominated by Galactic synchrotron emission. **b**, **c**, Residuals after fitting and removing only the foreground model (**b**) or the foreground and 21-cm models (**c**). **d**, Recovered model profile of the 21-cm absorption, with a signal-to-noise ratio of 37, amplitude of 0.53 K, centre frequency of 78.1 MHz and width of 18.7 MHz. **e**, Sum of the 21-cm model (**d**) and its residuals (**c**).

[1]School of Earth and Space Exploration, Arizona State University, Tempe, Arizona 85287, USA. [2]Haystack Observatory, Massachusetts Institute of Technology, Westford, Massachusetts 01886, USA. [3]Center for Astrophysics and Space Astronomy, University of Colorado, Boulder, Colorado 80309, USA. [4]Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Alonso de Ribera 2850, Concepción, Chile.

Table 1 | Sensitivity to possible calibration errors

Error source	Estimated uncertainty	Modelled error level	Recovered amplitude (K)
LNA S11 magnitude	0.1 dB	1.0 dB	0.51
LNA S11 phase (delay)	20 ps	100 ps	0.48
Antenna S11 magnitude	0.02 dB	0.2 dB	0.50
Antenna S11 phase (delay)	20 ps	100 ps	0.48
No loss correction	N/A	N/A	0.51
No beam correction	N/A	N/A	0.48

The estimated uncertainty for each case is based on empirical values from laboratory measurements and repeatability tests. Modelled error levels were chosen conservatively to be five and ten times larger than the estimated uncertainties for the phases and magnitudes, respectively. LNA, low-noise amplifier; S11, input reflection coefficient; N/A, not applicable.

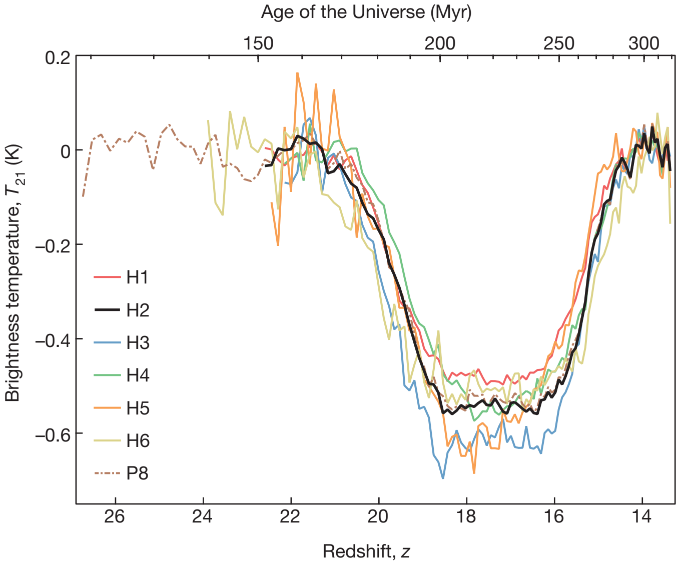


Figure 2 | Best-fitting 21-cm absorption profiles for each hardware case. Each profile for the brightness temperature T_{21} is added to its residuals and plotted against the redshift z and the corresponding age of the Universe. The thick black line is the model fit for the hardware and analysis configuration with the highest signal-to-noise ratio (equal to 52; H2; see Methods), processed using 60–99 MHz and a four-term polynomial (see equation (2) in Methods) for the foreground model. The thin solid lines are the best fits from each of the other hardware configurations (H1, H3–H6). The dash-dotted line (P8), which extends to $z > 26$, is reproduced from Fig. 1e and uses the same data as for the thick black line (H2), but a different foreground model and the full frequency band.

observations using restricted spectral bands yield nearly identical best-fitting absorption profiles, with the highest signal-to-noise ratio reaching 52. In Fig. 2 we show representative cases of these fits.

We performed numerous hardware and processing tests to validate the detection. The 21-cm absorption profile is observed in data that span nearly two years and can be extracted at all local solar times and at all local sidereal times. It is detected by two identically designed instruments operated at the same site and located 150 m apart, and even after several hardware modifications to the instruments, including orthogonal orientations of one of the antennas. Similar results for the absorption profile are obtained by using two independent processing pipelines, which we tested using simulated data. The profile is detected using data processed via two different calibration techniques: absolute calibration and an additional differencing-based post-calibration process that reduces some possible instrumental errors. It is also detected using several sets of calibration solutions derived from multiple laboratory measurements of the receivers and using multiple on-site measurements of the reflection coefficients of the antennas. We modelled the sensitivity of the detection to several possible calibration errors and in all cases recovered profile amplitudes that are within the reported confidence range, as summarized in Table 1. An EDGES high-band instrument operates between 90 MHz and 200 MHz at the same site using a nearly identical receiver and a scaled version of the low-band antennas. It does not produce a similar feature at the scaled frequencies[4]. Analysis of radio-frequency interference in the observations, including in the FM radio band, shows that the absorption profile is inconsistent with typical spectral contributions from these sources.

We are not aware of any alternative astronomical or atmospheric mechanisms that are capable of producing the observed profile. H II regions in the Galaxy have increasing optical depth with wavelength, blocking more background emission at lower frequencies, but they are observed primarily along the Galactic plane and generate monotonic spectral profiles at the observed frequencies. Radio-frequency recombination lines in the Galactic plane create a 'picket fence' of narrow absorption lines separated by approximately 0.5 MHz at the observed frequencies[5], but these lines are easy to identify and filter in the EDGES observations. The Earth's ionosphere weakly absorbs radio signals at the observed frequencies and emits thermal radiation from hot electrons, but models and observations imply a broadband effect that varies depending on the ionospheric conditions[6,7], including diurnal changes in the total electron content. This effect is fitted by our foreground model. Molecules of the hydroxyl radical and nitric oxide have spectral lines in the observed band and are present in the atmosphere, but the densities and line strengths are too low to produce substantial absorption.

The 21-cm line has a rest-frame frequency of 1,420 MHz. Expansion of the Universe redshifts the line to the observed band according to $\nu = 1{,}420/(1 + z)$ MHz, where z is the redshift, which maps uniquely to the age of the Universe. The observed absorption profile is the continuous superposition of lines from gas across the observed redshift range and cosmological volume; hence, the shape of the profile traces the history of the gas across cosmic time and is not the result of the properties of an individual cloud. The observed absorption profile is centred at $z \approx 17$ and spans approximately $20 > z > 15$.

The intensity of the observable 21-cm signal from the early Universe is given as a brightness temperature relative to the microwave background[8]:

$$T_{21}(z) \approx 0.023\ \mathrm{K} \times x_{\mathrm{HI}}(z)\left[\left(\frac{0.15}{\Omega_{\mathrm{m}}}\right)\left(\frac{1+z}{10}\right)\right]^{\frac{1}{2}}\left(\frac{\Omega_{\mathrm{b}}h}{0.02}\right)\left[1-\frac{T_{\mathrm{R}}(z)}{T_{\mathrm{S}}(z)}\right]$$

where x_{HI} is the fraction of neutral hydrogen, Ω_{m} and Ω_{b} are the matter and baryon densities, respectively, in units of the critical density for a flat universe, h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1}, T_{R} is the temperature of the background radiation, usually assumed to be from the background produced by the afterglow of the Big Bang, T_{S} is the 21-cm spin temperature that defines the relative population of the hyperfine energy levels, and the factor of 0.023 K comes from atomic-line physics and the average gas density. The spin temperature is affected by the absorption of microwave photons, which couples T_{S} to T_{R}, as well as by resonant scattering of Lyman-α photons and atomic collisions, both of which couple T_{S} to the kinetic temperature of the gas T_{G}.

The temperatures of the gas and the background radiation are coupled in the early Universe through Compton scattering. This coupling becomes ineffective in numerical models[9,10] at $z \approx 150$, after which primordial gas cools adiabatically. In the absence of stars or non-standard physics, the gas temperature is expected to be 9.3 K at $z = 20$, falling to 5.4 K at $z = 15$. The radiation temperature decreases more slowly owing to cosmological expansion, following $T_0(1 + z)$ with $T_0 = 2.725$, and so is 57.2 K and 43.6 K at the same redshifts, respectively. The spin temperature is initially coupled to the gas temperature as the gas cools below the radiation temperature, but eventually the decreasing density of the gas is insufficient to maintain this coupling and the spin temperature returns to the radiation temperature.

Over time, Lyman-α photons from early stars recouple the spin and gas temperatures[11], leading to the detected signal. The onset of the observed absorption profile at $z = 20$ places this epoch at an age of 180 million years, using the Planck 2015 cosmological parameters[12]. For the most extreme case, in which T_S is fully coupled to T_G, the gas and radiation temperatures calculated here yield a maximum absorption amplitude of 0.20 K at $z = 20$, which increases to 0.23 K at $z = 15$.

The presence of stars should eventually halt the cooling of the gas and ultimately heat it, because stellar radiation deposits energy into the gas and Lyman-line cooling has been modelled to be very small for the expected stellar properties[13]. As early stars die, they are expected to leave behind stellar remnants, such as black holes and neutron stars. The accretion disks around these remnants should generate X-rays, further heating the gas. At some point, the gas is expected to become hotter than the background radiation temperature, ending the absorption signal. The $z = 15$ edge of the observed profile places this transition around 270 million years after the Big Bang.

The ages derived for the events above fall within the range expected from many theoretical models[2]. However, the flattened shape of the observed absorption profile is uncommon in existing models, and could indicate that the initial flux of Lyman-α radiation from early stars was sufficiently large that the spin temperature saturated quickly to the gas temperature. We analysed models that assume a high Lyman-α flux at $z < 14$ using the EDGES high-band measurements and found a large fraction to be inconsistent with the data[4].

To produce the best-fitting profile amplitude of 0.5 K, the ratio T_R/T_S at the centre of the profile must be larger than 15, compared to a maximum of 7 allowed by the gas and radiation temperatures stated above. Even the lower confidence bound of 0.3 K for the observed profile amplitude is roughly 50% larger than the strongest predicted signal. For a standard gas temperature history, T_R would need to be more than 104 K to yield the best-fitting amplitude at the centre of the profile, whereas for a radiation temperature history given solely by the microwave background, T_G would need to be less than 3.2 K.

The observed profile amplitude could be explained if the gas and background radiation temperatures decoupled by $z \approx 250$ rather than $z \approx 150$, allowing the gas to begin cooling adiabatically earlier. A residual ionization fraction after the formation of atoms that is less than the expected fraction by nearly an order of magnitude would lead to sufficiently early decoupling. However, cross-validation between numerical models and their consistency with Planck observations suggest that the residual ionization fraction is already known to approximately 1% fractional accuracy.

Considering more exotic scenarios, interactions between dark matter and baryons can explain the observed profile amplitude (by lowering the gas temperature[14]) if the mass of dark-matter particles is below a few gigaelectronvolts and the interaction cross-section is greater than about 10^{-21} cm^2 (ref. 3). Existing models of other non-standard physics, including dark-matter decay and annihilation[15], accreting or evaporating primordial black holes[16], and primordial magnetic fields[17], all predict increased gas temperatures and are unlikely to account for the observed amplitude. It is possible that some of these sources could also increase T_R through mechanisms such as synchrotron emission associated with primordial black holes[18] or the relativistic electrons that result from the decay of metastable particles[19], but it is unclear whether this could compensate for the increased gas temperature. Measurements from ARCADE-2 (ref. 20) suggest an isotropic radio background that cannot be explained by known source populations, but this interpretation is still debated[21] and the unknown sources would need to be present at $z \approx 20$ to affect the radiation environment relevant to the observed signal.

Although we have performed many tests to be confident that the observed profile is from a global absorption of the microwave background by hydrogen gas in the early Universe, we still seek confirmation observations from other instruments. Several experiments similar to EDGES are underway. Those closest to achieving the performance required to verify the profile include the Large-Aperture Experiment to Detect the Dark Ages (LEDA)[22], the Sonda Cosmológica de las Islas para la Detección de Hidrógeno Neutro (SCI-HI)[23], the Probing Radio Intensity at high z from Marion (PRIZM) and the Shaped Antenna measurement of the background Radio Spectrum 2 (SARAS 2)[24]. The use of more sophisticated foreground models than we have used could lower the performance requirements of the hardware, and could lead to better profile recovery because low-order (smooth) modes in any fitted profile shape tend to be covariant with our foreground model and so are potentially under-constrained. Singular-value decomposition of training sets constructed from simulated instrument error terms and foreground contributions can produce optimized basis sets for model fitting[25,26]. We plan to apply these techniques to our data processing. The best measurement of the observed profile might ultimately be conducted in space, where the Earth's atmosphere and ionosphere cannot influence the propagation of the astronomical signal, potentially reducing the burden of the foreground model. The measurement could be made from the lunar farside[27]—either on the surface or from an orbit around the Moon—using the Moon as a shield to block FM radio signals and other Earth-based transmitters.

This result should bolster ongoing efforts to detect the statistical properties of spatial fluctuations in the 21-cm signal using interferometric arrays; it provides direct evidence that a signal exists for these telescopes to detect. The Hydrogen Epoch of Reionization Array[28] (HERA) will become operational over the next two years and will be able to characterize the power spectrum of redshifted 21-cm fluctuations between 100 MHz and 200 MHz during the epoch of reionization, when the 21-cm signal should be visible as emission above the background. The operational band of HERA is also expected to be extended to 50 MHz over this time. HERA will probably have sufficient thermal sensitivity to detect any power-spectrum signal associated with the EDGES observed profile and so might be able to validate the absorption signal reported here. But foregrounds have proven to be more challenging for interferometers than expected, and sufficient foreground mitigation to detect the 21-cm power spectrum has yet to be demonstrated by any of the currently operating arrays, including the Low-Frequency Array (LOFAR)[29] and the Murchison Widefield Array (MWA)[30]. Expansion of an existing Long Wavelength Array[31] station would provide the sensitivity necessary to pursue power-spectrum detection. When constructed, the Square Kilometre Array (SKA) Low-Frequency Aperture Array (https://www.skatelescope.org) should be able to detect the power spectrum associated with the absorption profile reported here and, eventually, to image the 21-cm signal.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 September 2017; accepted 24 January 2018.

1. Pritchard, J. R. & Loeb, A. 21 cm cosmology in the 21st century. *Rep. Prog. Phys.* **75,** 086901 (2012).
2. Cohen, A., Fialkov, A., Barkana, R. & Lotem, M. Charting the parameter space of the global 21-cm signal. *Mon. Not. R. Astron. Soc.* **472,** 1915–1931 (2017).
3. Barkana, R. Possible interaction between baryons and dark-matter particles revealed by the first stars. *Nature* **555,** https://doi.org/10.1038/nature25791 (2018).
4. Monsalve, R. A., Rogers, A. E. E., Bowman, J. D. & Mozdzen, T. J. Results from EDGES high-band. I. Constraints on phenomenological models for the global 21 cm signal. *Astrophys. J.* **847,** 64 (2017).
5. Alves, M. I. R. *et al.* The HIPASS survey of the Galactic plane in radio recombination lines. *Mon. Not. R. Astron. Soc.* **450,** 2025–2042 (2015).
6. Rogers, A. E. E., Bowman, J. D., Vierinen, J., Monsalve, R. A. & Mozdzen, T. Radiometric measurements of electron temperature and opacity of ionospheric perturbations. *Radio Sci.* **50,** 130–137 (2015).
7. Sokolowski, M. *et al.* The impact of the ionosphere on ground-based detection of the global epoch of reionization signal. *Astrophys. J.* **813,** 18 (2015).
8. Zaldarriaga, M., Furlanetto, S. R. & Hernquist, L. 21 centimeter fluctuations from cosmic gas at high redshifts. *Astrophys. J.* **608,** 622–635 (2004).
9. Ali-Haïmoud, Y. & Hirata, C. M. HyRec: a fast and highly accurate primordial hydrogen and helium recombination code. *Phys. Rev. D* **83,** 043513 (2011).

10. Shaw, J. R. & Chluba, J. Precise cosmological parameter estimation using COSMOREC. *Mon. Not. R. Astron. Soc*. **415,** 1343–1354 (2011).
11. Furlanetto, S. R. The global 21-centimeter background from high redshifts. *Mon. Not. R. Astron. Soc.* **371,** 867–878 (2006).
12. Planck Collaboration. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594,** A13 (2016).
13. Chen, X. & Miralda-Escude, J. The spin-kinetic temperature coupling and the heating rate due to Lyα scattering before reionization: predictions for 21 centimeter emission and absorption. *Astrophys. J.* **602,** 1–11 (2004).
14. Tashiro, H., Kadota, K. & Silk, J. Effects of dark matter-baryon scattering on redshifted 21 cm signals. *Phys. Rev. D* **90,** 083522 (2014).
15. Lopez-Honorez, L., Mena, O., Moliné, Á., Palomares-Ruiz, S. & Vincent, A. C. The 21 cm signal and the interplay between dark matter annihilations and astrophysical processes. *J. Cosmol. Astropart. Phys.* **8,** 4 (2016).
16. Tashiro, H. & Sugiyama, N. The effect of primordial black holes on 21-cm fluctuations. *Mon. Not. R. Astron. Soc*. **435,** 3001–3008 (2013).
17. Schleicher, D. R. G., Banerjee, R. & Klessen, R. S. Influence of primordial magnetic fields on 21 cm emission. *Astrophys. J.* **692,** 236–245 (2009).
18. Biermann, P. L. *et al.* Cosmic backgrounds due to the formation of the first generation of supermassive black holes. *Mon. Not. R. Astron. Soc*. **441,** 1147–1156 (2014).
19. Cline, J. M. & Vincent, A. C. Cosmological origin of anomalous radio background. *J. Cosmol. Astropart. Phys.* **2,** 11 (2013).
20. Seiffert, M. *et al.* Interpretation of the Arcade 2 absolute sky brightness measurement. *Astrophys. J.* **734,** 6 (2011).
21. Vernstrom, T., Norris, R. P., Scott, D. & Wall, J. V. The deep diffuse extragalactic radio sky at 1.75 GHz. *Mon. Not. R. Astron. Soc*. **447,** 2243–2260 (2015).
22. Bernardi, G. *et al.* Bayesian constraints on the global 21-cm signal from the cosmic dawn. *Mon. Not. R. Astron. Soc.* **461,** 2847–2855 (2016).
23. Voytek, T. C., Natarajan, A., Jáuregui García, J. M., Peterson, J. B. & López-Cruz, O. Probing the dark ages at $z \sim 20$: the SCI-HI 21 cm all-sky spectrum experiment. *Astrophys. J.* **782,** L9 (2014).
24. Singh, S. *et al.* First results on the epoch of reionization from first light with SARAS 2. *Astrophys. J.* **845,** L12 (2017).
25. Switzer, E. R. & Liu, A. Erasing the variable: empirical foreground discovery for global 21 cm spectrum experiments. *Astrophys. J.* **793,** 102 (2014).
26. Vedantham, H. K. *et al.* Chromatic effects in the 21 cm global signal from the cosmic dawn. *Mon. Not. R. Astron. Soc*. **437,** 1056–1069 (2014).
27. Burns, J. O. *et al.* A space-based observational strategy for characterizing the first stars and galaxies using the redshifted 21 cm global spectrum. *Astrophys. J.* **844,** 33 (2017).
28. DeBoer, D. R. *et al.* Hydrogen Epoch of Reionization Array (HERA). *Publ. Astron. Soc. Pac.* **129,** 045001 (2017).
29. Patil, A. H. *et al.* Upper limits on the 21 cm epoch of reionization power spectrum from one night with LOFAR. *Astrophys. J.* **838,** 65 (2017).
30. Beardsley, A. P. *et al.* First season MWA EoR power spectrum results at redshift 7. *Astrophys. J.* **833,** 102 (2016).
31. Dowell, J., Taylor, G. B., Schinzel, F. K., Kassim, N. E. & Stovall, K. The LWA1 low frequency sky survey. *Mon. Not. R. Astron. Soc*. **469,** 4537–4550 (2017).

Acknowledgements We thank CSIRO for providing site infrastructure and access to facilities. We thank the MRO Support Facility team, especially M. Reay, L. Puls, J. Morris, S. Jackson, B. Hiscock and K. Ferguson. We thank C. Bowman, D. Cele, C. Eckert, L. Johnson, M. Goodrich, H. Mani, J. Traffie and K. Wilson for instrument contributions. We thank R. Barkana for theory contributions and G. Holder, T. Vachaspati, C. Hirata and J. Chluba for exchanges. We thank C. Lonsdale, H. Johnson, J. Hewitt and J. Burns. We thank C. Halleen and M. Halleen for site and logistical support. This work was supported by the NSF through awards AST-0905990, AST-1207761 and AST-1609450. R.A.M. acknowledges support from the NASA Ames Research Center (NNX16AF59G) and the NASA Solar System Exploration Research Virtual Institute (80ARC017M0006). This work makes use of the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamatji people as the traditional owners of the site of the observatory.

Author Contributions J.D.B., R.A.M. and A.E.E.R. contributed to all activities. N.M. and T.J.M. modelled instrument properties, performed laboratory calibrations and contributed to the preparation of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.D.B. (judd.bowman@asu.edu).

Reviewer Information *Nature* thanks S. Weinreb and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Instrument. The EDGES experiment is located at the Murchison Radio-astronomy Observatory (MRO) in Western Australia (26.72° S, 116.61° E), which is the same radio-quiet[32] site used by the Australian SKA precursor, the Murchison Widefield Array and the planned SKA Low-Frequency Aperture Array. An early version of EDGES placed an empirical lower limit on the duration of the epoch of reionization[33]. EDGES currently consists of three instruments: a high-band instrument[4,34,35] that is sensitive to 90–200 MHz ($14 > z > 6$) and two low-band instruments (low-1 and low-2) that operate over the range 50–100 MHz ($27 > z > 13$). Each instrument yields spectra with 6.1-kHz resolution. In each instrument, sky radiation is collected by a wideband dipole-like antenna that consists of two rectangular metal panels mounted horizontally above a metal ground plane. Similar compact dipole antennas are used elsewhere in radio astronomy[36,37]. A receiver is installed underneath the ground plane and a balun[38] is used to guide radiation from the antenna panels to the receiver. A rectangular shroud surrounds the base of the balun to shield the vertical currents in the balun tubes, which are strongest at the base. This lowers the gain towards the horizon owing to the vertical currents.

Extended Data Fig. 1 shows a block diagram of the system. A mechanical input switch at the front of the receiver enables the antenna to be connected to a remote vector network analyser (VNA) to measure the antenna reflection coefficients accurately or to be connected to the primary receiver path to measure the sky noise spectrum. When measuring the sky noise spectrum a second mechanical switch connects the low-noise amplifier (LNA) to the antenna or to a 26-dB attenuator, which acts as a load or a well matched noise source, depending on the state of the electrical switch on the noise source. This performs a three-position switching operation[39] between ambient and 'hot' internal noise sources and the antenna, which is needed to provide the first stage of processing discussed below. After the LNA and post-amplifier, another noise source is used to inject noise at less than 45 MHz. This 'out of band' conditioning improves the linearity and dynamic range of the analogue-to-digital converter (ADC), which are needed for the accurate cancellation of the receiver bandpass that is afforded by the three-position switching. A thermoelectric system maintains a constant temperature in the receiver in the field and in the laboratory. The system mitigates against radio-frequency interference (RFI) using designs and analysis strategies adapted from the Deuterium Array[40]. Similar approaches to instrument design are used for SARAS 2 (ref. 41) and LEDA[42].

The design of the low-band instrument differs from the published descriptions of the design of the high-band instrument only by the following. First, a 3-dB attenuator is added within the LNA before the pseudomorphic high-electron-mobility transistor at the input. The attenuator improves the LNA impedance match and thereby reduces the sensitivity to measurement errors of the LNA and antenna reflection coefficients, especially to errors in the reflection phase, while adding only a small amount of noise compared with the sky noise. Larger values of attenuation would begin to add substantial noise at 100 MHz. Second, a scaled antenna is used that is precisely double the size of the antenna of the high-band instrument. Third, a larger ground plane is used. Each ground plane of the low-band instrument consists of a 2 m × 2 m solid metal central assembly surrounded by metal mesh that spans 30 m × 30 m, with the outer 5 m shaped as saw-tooth perforated edges. Low-1 was initially operated with a 10 m × 10 m ground plane and later extended to full size. The full-size 30 m × 30 m ground plane reduces the chromaticity of the beam and makes it less sensitive to conditions of the soil. Extended Data Fig. 2 shows the low-1 and low-2 antennas, Extended Data Fig. 3 the measured reflection coefficients and Extended Data Fig. 4 cuts through the model of the antenna beam pattern.

Calibration. We implement end-to-end absolute calibration for the low-band instruments following the techniques developed for the high-band instrument[34,43]. The calibration procedure involves taking reference spectra in the laboratory with the receiver connected to hot and ambient loads and to open and shorted cables. Similar techniques are used in other microwave measurements[44,45]. Reflection coefficient measurements using a VNA are acquired for the calibration sources and the LNA. The input connection to the receiver box provides the reference plane for all VNA measurements. To correct for the losses in the hot load used in the laboratory for calibration, full scattering parameters are measured for the short cable from the heated resistor in the hot load. The accuracy of scattering parameter measurements is improved by accounting for the actual resistance of the 50-Ω load used for VNA calibration and the added inductance in the load due to skin effects[46] in the few millimetres of transmission line between its internal termination and the reference plane[47].

The calibration spectra and reflection coefficient measurements acquired in the laboratory are used to solve for the free parameters[34] in the equations that account for the impedance mismatches between the receiver and the antenna and for the correlated and uncorrelated LNA noise waves. Laboratory calibration is performed with the receiver temperature controlled to the default 25 °C, and at 15 °C and 35 °C to assess the thermal dependence of the calibration parameters. Extended Data Fig. 5 shows calibration parameter solutions for both receivers.

Following calibration in the laboratory, a test is performed by measuring the spectrum of an approximately 300 K passive load with deliberate impedance mismatch that approximately mimics the reflection from the antenna in magnitude and phase. We call this device an 'artificial antenna simulator'. The reflection coefficient of the antenna simulator is measured and applied to yield calibrated integrated spectra. Extended Data Fig. 3 shows measured reflection coefficients for the artificial antenna simulator. Once corrected, the integrated spectra are expected to be spectrally flat, with a noise temperature that matches the physical temperature of the passive load. The flatness of an integrated spectrum is quantified through the r.m.s. of the residuals after subtracting a constant term. The typical r.m.s. of the residuals is 0.025 K over the range 50–100 MHz. If a three-term polynomial (equation (2) with $N = 3$) is subtracted, then the residuals decrease to about 0.015 K and are limited by integration time.

A second test of the calibration is performed by measuring the spectrum of a noise source followed by a filter and an approximately 3-m cable that adds about 30 ns of two-way delay. The device yields a spectrum that is similar in shape to the sky foreground with a strength of about 10,000 K (seven times larger than the typical sky temperature observed by EDGES) at 75 MHz and has a reflection coefficient of −6 dB in magnitude with a phase slope similar to that of the antenna. Typical residuals are less than 300 mK after subtracting a five-term polynomial (equation (2)) and are limited by integration time. Assuming any residuals scale with input power, this value corresponds to residuals of 45 mK at the typical observed sky temperature. This test is more sensitive than the passive simulator, especially to errors in the measurements of the reflection coefficient, because the signal is 33 times stronger than that of the approximately 300-K load and because the magnitude of the reflection of this simulator is larger than that of the passive simulator and of the real antenna.

Losses in the balun and losses due to the finite ground plane are corrected for during data processing using models. The balun-loss model is validated against scattering parameter measurements. Frequency-dependent beam effects are compensated for by modelling and subtracting the spectral structure using electromagnetic beam models and a diffuse sky map template[35]. The nominal beam model accounts for the finite metal ground plane over soil with a relative permittivity of 3.5 and conductivity of 2×10^{-2} S m^{-1} (ref. 48). The sky template is produced by extrapolating the 408-MHz all-sky radio map[49] to the observed frequencies using a spectral index in brightness temperature of −2.5 (refs 35, 43).

Data and processing. Examples of raw and processed data are shown in Extended Data Fig. 6. Data processing involves three primary stages. In the first stage, three raw spectra that have each been accumulated for 13 s from the antenna input and two internal reference noise sources are converted into a single partially calibrated spectrum[39]. Individual 6.1-kHz channels above a fixed power threshold are assigned zero weight to remove RFI. The threshold is normally set at three times the r.m.s. of the residuals after substacting a constant and a slope in a sliding 256-spectral-channel window. Similarly, any partially calibrated spectrum with an average power above that expected from the sky or with large residuals is discarded. A weighted average of many successive spectra is taken, typically over several hours. Outlier channels after a Fourier series fit to the entire accumulated spectrum are again assigned zero weight. This second pass assigns zero weight to lower levels of RFI and broader RFI signals than does the initial pass.

In the second stage of processing, the partially calibrated spectra are fully calibrated using the calibration parameters from the laboratory and the antenna reflection coefficient measurements taken periodically in the field. Beam-chromaticity corrections are applied after averaging the model over the same range of local sidereal times as for the spectra. The spectra are then corrected for the balun and ground-plane loss and output with a typical smoothing to spectral bins with 390.6-kHz resolution.

In the third stage of processing, spectra for each block of local sidereal time of several hours within each day are fitted with a foreground model (see below for description of models). An r.m.s. value of the residuals is computed for each block and blocks above a selected threshold are discarded, typically because of broadband RFI—or solar activity in daytime data—that was not detected in the earlier processing stages. A weighted average is then taken of the accepted blocks and a weighted least-squares solution is determined using a foreground model along with the model that represents the 21-cm absorption signal. Extended Data Fig. 7 shows the final weights for each spectral bin, equivalent to the RFI occupancy.

The observations used for the primary analysis presented here are from low-1, spanning 2016 day 252 through to 2017 day 94 (configuration H2 below). The data are filtered to retain only local Galactic hour angles (GHA) of 6–18 h; GHA is equivalent to the local sidereal time offset by 17.75 h.

Parameter estimation. The polynomial foreground model used for the analysis presented in Fig. 1 is physically motivated, with five terms based on the known spectral properties of the Galactic synchrotron spectrum and Earth's ionosphere[6,50]:

$$T_{\mathrm{F}}(\nu) \approx a_0\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2.5} + a_1\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2.5}\log\left(\frac{\nu}{\nu_{\mathrm{c}}}\right) + a_2\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2.5}\left[\log\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)\right]^2 + a_3\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-4.5} + a_4\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2} \quad (1)$$

where $T_{\mathrm{F}}(\nu)$ is the brightness temperature of the foreground emission, ν is the frequency, ν_{c} is the centre frequency of the observed band and the coefficients a_n are fitted to the data. The above function is a linear approximation, centred on ν_{c}, to

$$T_{\mathrm{F}}(\nu) = b_0\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2.5+b_1+b_2\log(\nu/\nu_{\mathrm{c}})} \mathrm{e}^{-b_3(\nu/\nu_{\mathrm{c}})^{-2}} + b_4\left(\frac{\nu}{\nu_{\mathrm{c}}}\right)^{-2}$$

which is connected directly to the physics of the foreground and the ionosphere. The factor of −2.5 in the first exponent is the typical power-law spectral index of the foreground, b_0 is an overall foreground scale factor, b_1 allows for a correction to the typical spectral index of the foreground (which varies by roughly 0.1 across the sky) and b_2 captures any contributions from a higher-order foreground spectral term[51,52]. Ionospheric contributions are contained in b_3 and b_4, which allow for the ionospheric absorption of the foreground and emission from hot electrons in the ionosphere, respectively. This model can also partially capture some instrumental effects, such as additional spectral structure from chromatic beams or small errors in calibration.

We also use a more general polynomial model in many of our trials that enables us to explore signal recovery with varying numbers of polynomial terms:

$$T_{\mathrm{F}}(\nu) = \sum_{n=0}^{N-1} a_n \nu^{n-2.5} \quad (2)$$

where N is the number of terms and the coefficients a_n are again fitted to the data. As with the physical model, the −2.5 in the exponent makes it easier for the model to match the foreground spectrum. Both foreground models yield consistent absorption profile results.

The 21-cm absorption profile is modelled as a flattened Gaussian:

$$T_{21}(\nu) = -A\left(\frac{1-\mathrm{e}^{-\tau \mathrm{e}^{B}}}{1-\mathrm{e}^{-\tau}}\right)$$

where

$$B = \frac{4(\nu-\nu_0)^2}{w^2}\log\left[-\frac{1}{\tau}\log\left(\frac{1+\mathrm{e}^{-\tau}}{2}\right)\right]$$

and A is the absorption amplitude, ν_0 is the centre frequency, w is the full-width at half-maximum and τ is a flattening factor. This model is not a description of the physics that creates the 21-cm absorption profile, but rather is a suitable functional form to capture the basic shape of the profile. Extended Data Fig. 8 shows the best-fitting profile and residuals from fits by the two foreground models.

We report parameter fits from a gridded search over the parameters ν_0, w and τ in the 21-cm model. For each step in the grid, we conduct a linear weighted least-squares fit, solving simultaneously for the foreground coefficients and the amplitude of the absorption profile. The best-fitting absorption profile maximizes the signal-to-noise ratio in the gridded search. The uncertainty in the amplitude fit accounts for covariance between the foreground coefficients and the amplitude of the profile and for noise.

Fitting both foreground and 21-cm models simultaneously yields residuals that decrease with integration time with an approximately noise-like ($1/\sqrt{t}$) trend for the duration of the observation, whereas fitting only the foreground model yields residuals that decrease with time for the first approximately 10% of the integration and then saturate, as shown in Extended Data Fig. 9.

We also performed a Markov chain Monte Carlo (MCMC) analysis (Extended Data Fig. 10) for the H2 case using a five-term polynomial (equation (2)) for the foreground model and a subset of the band covering 60–94 MHz. The amplitude parameter is most covariant with the flattening. The 99% statistical confidence intervals on the four 21-cm model parameters are: $A = 0.52^{+0.42}_{-0.18}$ K, $\nu_0 = 78.3^{+0.2}_{-0.3}$ MHz, $w = 20.7^{+0.8}_{-0.7}$ MHz and $\tau = 6.5^{+5.6}_{-2.5}$. These intervals do not include any systematic error from differences across the hardware configurations and processing trials. When the flattening parameter is fixed to $\tau = 7$, statistical uncertainty in the 21-cm model amplitude fit is reduced to approximately ±0.02 K. Extended Data Table 1 shows that the various hardware configurations and processing trials with fixed $\tau = 7$ yield best-fitting parameter ranges of $0.37\,\mathrm{K} < A < 0.67\,\mathrm{K}$, $77.4\,\mathrm{MHz} < \nu_0 < 78.5\,\mathrm{MHz}$ and $17.0\,\mathrm{MHz} < w < 22.8\,\mathrm{MHz}$. This systematic variation is probably due to the limited data in the some of the configurations, small calibration errors and residual chromatic beam effects, and potentially to structure in the Galactic foreground that increases when the Galactic plane is overhead. For each parameter, taking the outer bounds of the statistical confidence ranges from the comprehensive MCMC analysis for H2 and the best-fitting variations between validation trials in Extended Data Table 1 yields the estimate of the 99% confidence intervals that we report in the main text.

Verification tests. Here we list the tests that we performed to verify the detection. The absorption profile is detected from data obtained in the following hardware configurations: H1, low-1 with 10 m × 10 m ground plane; H2, low-1 with 30 m × 30 m ground plane; H3, low-1 with 30 m × 30 m ground plane and recalibrated receiver; H4, low-2 with north–south dipole orientation; H5, low-2 with east–west dipole orientation; and H6, low-2 with east–west dipole orientation and the balun shield removed to check for any resonance that might result from a slot antenna being formed in the joint between the two halves of the shield.

The absorption profile is detected in data processed with the following configurations: P1, all hardware cases, using two independent processing pipelines; P2, all hardware cases, divided into temporal subsets; P3, all hardware cases, with chromatic beam corrections on or off; P4, all hardware cases, with ground loss and balun loss corrections on or off; P5, all hardware cases, calibrated with four different antenna reflection coefficient measurements; P6, all hardware cases, using a four-term foreground model (equation (2)) over the frequency range 60–99 MHz; P7, all hardware cases, using a five-term foreground model (equation (2)) over the frequency range 60–99 MHz; P8, all hardware cases, using the physical foreground model (equation (1)) over the frequency range 51–99 MHz; P9, all hardware configurations, using various additional combinations of four-, five-, and six-term foreground models (equation (2)) and various frequency ranges; P10, H2 binned by local sidereal time/GHA; P11, H2 binned by UTC; P12, H2 binned by buried conduit temperature as a proxy for the ambient temperature at the receiver and the temperature of the cable that connects the receiver frontend under the antenna to the backend in the control hut; P13, H2 binned by Sun above or below the horizon; P14, H2 binned by Moon above or below the horizon; P15, H2 with added post-calibration calculated by subtracting scaled Galaxy-up spectra from Galaxy-down spectra; P16, H2 calibrated with low-2 solutions; P17, H4 calibrated with laboratory measurements at 15 °C and 35 °C; and P18, H2 and H3 calibrated with laboratory measurements spanning two years.

Extended Data Table 1 lists the properties of the profile for each of the hardware configurations (H1–H6) with the standard processing configuration (P6); Fig. 2 illustrates the corresponding best-fitting profiles. The variations in signal-to-noise ratio between the configurations are largely explained by differences in the total integration time for each configuration, except for H1, which was limited by its ground-plane performance. We acquired the most data in configuration H1, with approximately 11 months of observations, followed by H2 with 6 months. The other configurations were each operated for 1–2 months before the analysis presented here was performed. Extended Data Table 2 lists the profile amplitudes for data binned by GHA for both processing pipelines used in configuration P1.

The following additional verification tests were performed to check specific aspects of the instrument, laboratory calibration and processing pipelines. (1) We processed simulated data and recovered injected profiles. (2) We searched for a similar profile at the scaled frequencies in high-band data and found no corresponding profile. (3) We measured the antenna reflection coefficients of low-2 with the VNA connected to its receiver via a short 2-m cable and found nearly identical results to when the 100-m cable was used (under normal operation). (4) We acquired *in situ* reflection coefficient measurements that matched our model predictions of the low-2 balun with the antenna terminal shorted and open; this was done to verify our model for the balun loss. (5) We tested the performance of the receivers in the laboratory using artificial antenna sources connected to the receivers directly, as described above. (6) We cross-checked our beam models using three electromagnetic numerical solvers: CST, FEKO and HFSS. Although no beam model is required to detect the profile because the EDGES antenna is designed to be largely achromatic, we performed the cross-check because we apply beam corrections in the primary analysis.

Sensitivity to systematic errors. Here we discuss in more detail several primary categories of potential systematic errors and the validation steps that we performed. *Beam and sky effects.* Beam chromaticity is larger than can be accounted for with electromagnetic models of the antenna on an infinite ground plane[53]. For both ground plane sizes, the r.m.s. of the residuals to low-order foreground polynomial model fits of the data matched electromagnetic modelling when the model accounted for the finite ground-plane size and included the effects of the dielectric constant and the conductivity of the soil under the ground plane. The residual structures themselves matched qualitatively.

Comparing electromagnetic solvers for beam models, we found that for infinite ground-plane models the change in the absolute gain of the beam with frequency at every viewing angle (θ, ϕ) was within ±0.006 between solvers and that residuals after foreground fits to simulated spectra were within a factor of two. For models with finite ground planes and real soil properties, we found that correcting H1 data using beam models from FEKO and HFSS in integral solver modes resulted in nearly identical 21-cm model parameter values, although using an HFSS model for the larger ground plane in H2 resulted in a fit to the profile that had a lower signal-to-noise ratio than when using the FEKO model, but still a higher signal-to-noise ratio than when no beam correction was used (see Extended Data Table 1).

The low-2 instrument was deployed 100 m west of the control hut, whereas the low-1 instrument was 50 m east of the hut. In the east–west antenna orientation, the low-2 dipole-response null was aimed approximately at the control hut and the beam pattern on the sky was rotated relative to north–south. Obtaining consistent absorption profiles with the two sizes of low-1 ground planes (H1 versus H2 or H3) and with both low-2 antenna orientations (H4 versus H5 or H6) suggests that beam effects are not responsible for the profile, while obtaining the same results from both low-2 antenna orientations also disfavours polarized sky emission as a possible source of the profile. Obtaining consistent absorption profiles with the low-1 and low-2 instruments at different distances from the control hut and with both low-2 antenna orientations suggests that it is unlikely that the observed profile is produced by reflections of sky noise from the control hut or other surrounding objects or caused by RFI from the hut. Our understanding of hut reflections is further validated by the appearance of small sinusoidal ripples after subtracting a nine-term foreground model (equation (2)) from the low-1 spectra at GHA 20. These ripples are consistent with models of hut reflections and not evident at other GHAs or in the low-2 data.

Gain and loss errors. Many possible instrumental systematic errors and atmospheric effects that could potentially mimic the observed absorption profile are due to inaccurate or unaccounted for gains or losses in the propagation path within the instrument or Earth's atmosphere. If present, these effects would be proportional to the total sky noise power entering the system. The total sky noise power received by EDGES varies by a factor of three over GHA. If the observed absorption profile were due to gain or loss errors, the amplitude of the profile would be expected to vary in GHA proportional to the sky noise.

We tested for these errors by fitting for the absorption profile in observations binned by GHA in 4-h and 6-h blocks using H2 data. The test is complicated by the increase in chromatic-beam effects in spectra when the sky noise power is large owing to the presence of the Galactic plane in the antenna beam. We compensated for this by increasing the foreground model to up to six polynomial terms for the GHA analysis and using the FEKO antenna beam model to correct for beam effects. As evident in Extended Data Table 2, the best-fitting amplitudes averaged over each GHA bin are consistent within the reported uncertainties and exhibit no substantial correlation with sky noise power. The same test performed using a four-term polynomial foreground model (equation (2)) did yield variations with GHA, as did tests performed on data from low-1 with the 10 m × 10 m ground plane. We attribute the failure of these two cases to beam effects and possible foreground structure. Other cases tested had insufficient data for conclusive results, but did not show any correlation with the total sky noise power.

The artificial antenna measurements described in Methods section 'Calibration' provide verification of the smooth passband of the receiver after calibration. Because we observe the 0.5-K signature for all foreground conditions, including low foregrounds of about 1,500 K at about 78 MHz, if the observed profile were an instrumental artefact due to an error in gain of the receiver, then we would expect to see a scaled version of the profile with an amplitude of $0.5\,\mathrm{K} \times (300\,\mathrm{K}/1{,}500\,\mathrm{K}) = 0.1\,\mathrm{K}$ when measuring the approximately 300-K artificial antenna. Instead, we see a smooth spectrum structure at the approximately 0.025-K level. With the 10,000-K artificial antenna, we would expect to see a 3.3-K profile that yields 0.5-K residuals after subtracting a five-term polynomial fit (equation (2)); instead we find residuals that are less than 0.3 K.

Receiver calibration errors are disfavoured as the source of the observed profile. Three verification tests were performed to investigate this possibility specifically by processing data with inaccurate calibration parameter solutions. First, in verification test P18 we processed H2 and H3 datasets using each of the three low-1 receiver calibration solutions shown in Extended Data Fig. 5a–g. The observed profile was detected in each case, indicating that the detection is robust to these small drifts in the calibration parameters over the two-year period spanning the use of the low-1 receiver. Second, in verification test P17 we processed H4 observations using the calibration solutions derived from laboratory measurements acquired with the receiver temperature held at both 15 °C and 35 °C, even though it was controlled to 25 °C for all observations. The profile was recovered even for these larger calibration differences and so we infer that the detection is robust to the much smaller variations of typically 0.1 °C in the receiver temperature around its set point during operation. Third, as a final check of the receiver properties, in P16 we calibrated the H2 dataset from low-1 using the receiver calibration solutions derived for low-2. The profile was recovered using a seven-term foreground model (equation (2)) over the range 53–99 MHz. This provides evidence that both receivers have generally similar properties and spectrally smooth responses; otherwise, we would not expect the calibration solutions to be interchangeable in this manner.

RFI and FM radio. RFI is found to be minimal in EDGES low-band measurements. We rule out locally generated broadband RFI from the control hut and a nearby ASKAP dish (more than 150 m away) as the source of the profile because of the consistent profiles observed by both instruments and both low-2 antenna orientations, as noted above. There are no licensed digital TV transmitters in Australia below 174 MHz (https://www.acma.gov.au, ITU RCC-06). We analysed observations and rule out the FM radio band, which spans 87.5–108 MHz, as the cause of the high-frequency edge of the observed profile. FM transmissions within about 3,000 km of the MRO could be scattered from aeroplanes or from meteors that burn up at an altitude of about 100 km in the mesosphere. Inspection of channels removed by our RFI detection algorithms and of spectral residuals using the instrument's raw 6.1-kHz resolution, which oversamples the minimum 50-kHz spacing of the FM channel centres, shows that these signals are sparse and transient and show up after removal as mostly zero-weighted channels. More persistent worldwide FM signals reflected from the Moon have been measured[54] from the MRO with flux density about 100 Jy. We find evidence for a sharp step of around 0.05 K at 87.5 MHz in our binned spectra when the Moon is above 45° elevation, which can be eliminated by using only data from when the Moon is below the horizon.

Atmospheric molecular lines. Atmospheric nitrous oxide line absorption was modelled using a line strength of $10^{-12.7}$ nm^2 MHz at 300 K based on values from the spectral line catalogue maintained by the Molecular Spectroscopy Team at the Jet Propulsion Laboratory and an abundance of 70 parts per billion. We assumed a 3,000-K sky-noise temperature and a line-of-sight path through the atmosphere at 8° elevation and integrated over the altitude range 10–120 km. We find up to 0.001 mK of absorption per line. With approximately 100 individual lines between 50 MHz and 100 MHz, we conservatively estimate a maximum possible contribution of 0.1 mK.

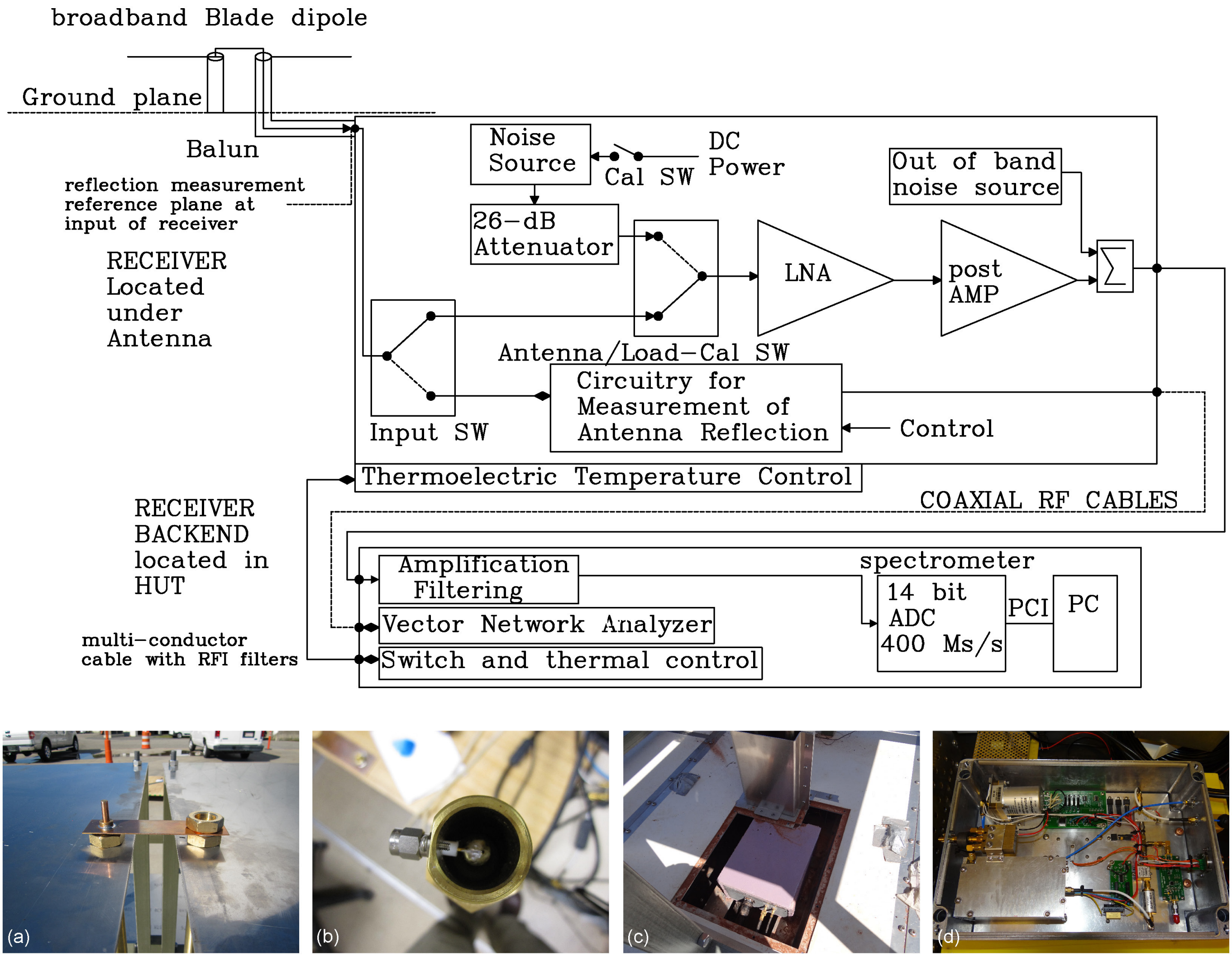
Gas temperature and residual ionization fraction. For the gas thermal calculations, we used CosmoRec[10] to model the evolution of the electron temperature and residual ionization fraction for $z < 3{,}000$. We verified the output against solutions to equations[55] for the dominant contributions to the electron temperature evolution of adiabatic expansion and Compton scattering. We assume that the gas temperature is in equilibrium with the electron temperature. To determine the residual ionization fraction that is required to produce sufficiently cold gas to account for the observed profile amplitude, we modelled a partial ionization step function in redshift. We used the ionization fraction from CosmoRec for redshifts above the transition and a constant final ionization fraction below the transition. We performed a grid search in transition redshift and final ionization fraction to identify the lowest transition redshift for the largest final ionization fraction that results in the required gas temperature. We found that a final ionization fraction of around 3×10^{-5} reached by $z \approx 500$ would be sufficient to produce the required gas temperature. This is nearly an order of magnitude lower than the expected ionization fraction of around 2×10^{-4} at similar ages from CosmoRec.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

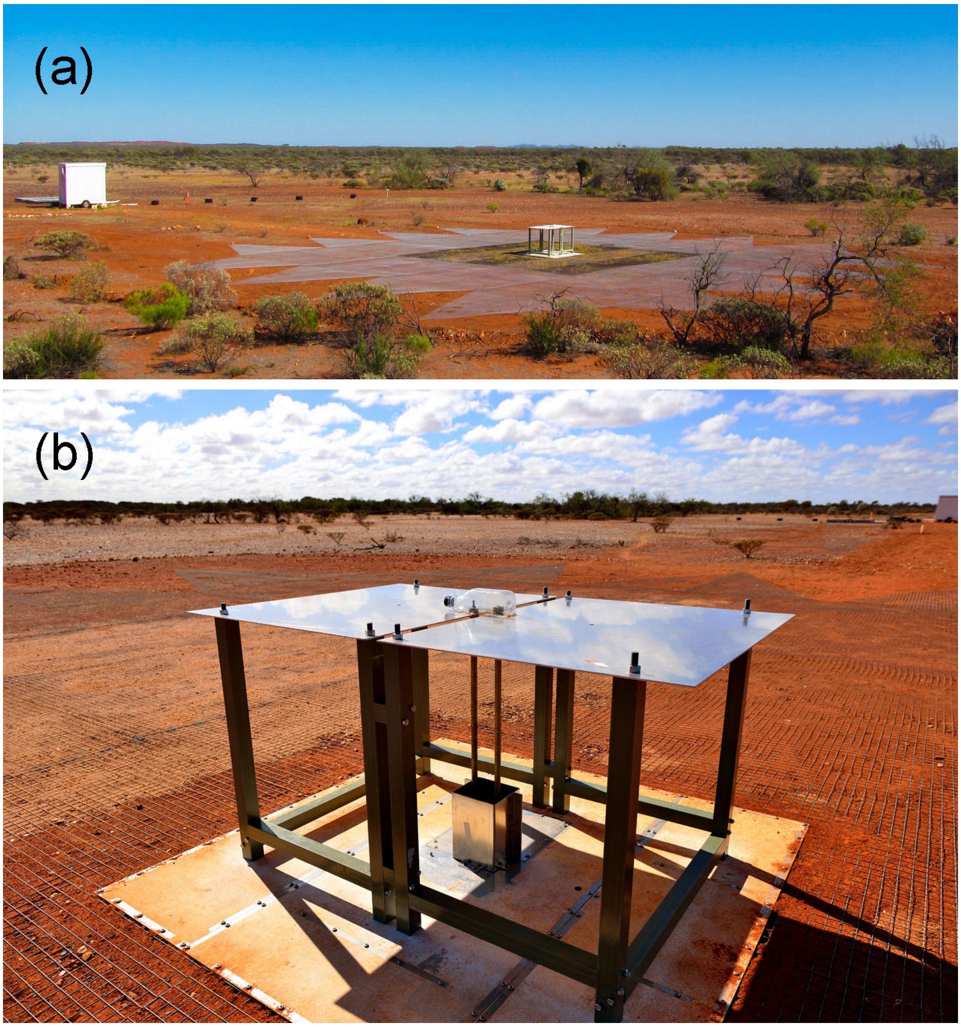
Code availability. The code that supports the findings of this study is available from the corresponding author on reasonable request.

32. Bowman, J. & Rogers, A. E. E. VHF-band RFI in geographically remote areas. In *Proc. RFI Mitigation Workshop* id.30, https://pos.sissa.it/107/030/pdf (Proceedings of Science, 2010).
33. Bowman, J. D. & Rogers, A. E. E. Lower Limit of Δz>0.06 for the duration of the reionization epoch. *Nature* **468,** 796–798 (2010).
34. Monsalve, R. A., Rogers, A. E. E., Bowman, J. D. & Mozdzen, T. J. Calibration of the EDGES high-band receiver to observe the global 21 cm signature from the epoch of reionization. *Astrophys. J.* **835,** 49 (2017).
35. Mozdzen, T. J., Bowman, J. D., Monsalve, R. A. & Rogers, A. E. E. Improved measurement of the spectral index of the diffuse radio background between 90 and 190 MHz. *Mon. Not. R. Astron. Soc.* **464,** 4995–5002 (2017).
36. Raghunathan, A., Shankar, N. U. & Subrahmanyan, R. An octave bandwidth frequency independent dipole antenna. *IEEE Trans. Anenn. Propag.* **61,** 3411–3419 (2013).
37. Ellingson, S. W. Antennas for the next generation of low-frequency radio telescopes. *IEEE Trans. Antenn. Propag.* **53,** 2480–2489 (2005).
38. Roberts, W. K. A new wide-band balun. *Proc. IRE* **45,** 1628–1631 (1957).

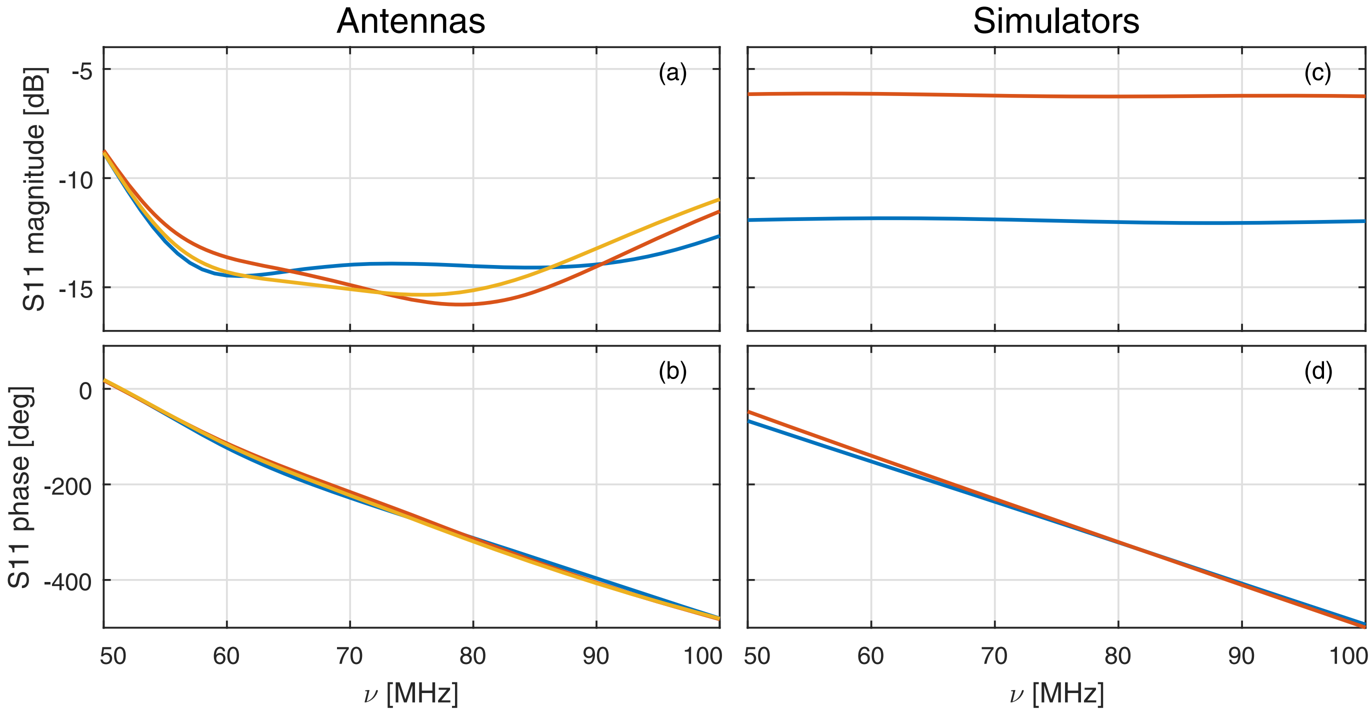
39. Bowman, J. D., Rogers, A. E. E. & Hewitt, J. N. Toward empirical constraints on the global redshifted 21 cm brightness temperature during the epoch of reionization. *Astrophys. J.* **676,** 1–9 (2008).
40. Rogers, A. E. E., Pratap, P., Carter, J. C. & Diaz, M. A. Radio frequency interference shielding and mitigation techniques for a sensitive search for the 327 MHz line of deuterium. *Radio Sci.* **40,** RS5S17 (2005).
41. Singh, S. *et al.* SARAS 2: a spectral radiometer for probing cosmic dawn and the epoch of reionization through detection of the global 21 cm signal. Preprint at https://arxiv.org/abs/1710.01101 (2017).
42. Price, D. C. *et al.* Design and characterization of the Large-Aperture Experiment to Detect the Dark Age (LEDA) radiometer systems. Preprint at https://arxiv.org/abs/1709.09313 (2017).
43. Rogers, A. E. E. & Bowman, J. D. Absolute calibration of a wideband antenna and spectrometer for accurate sky noise temperature measurement. *Radio Sci*. **47,** RS0K06 (2012).
44. Hu, R. & Weinreb, S. A novel wide-band noise-parameter measurement. *IEEE Trans. Microw. Theory Tech.* **52,** 1498–1507 (2004).
45. Belostotski, L. A calibration method for RF and microwave noise sources. *IEEE Trans. Microw. Theory Tech.* **59,** 178–187 (2011).
46. Ramo, S. & Whinnery, J. R. *Fields and Waves in Modern Radio* Ch. 6 (Wiley, 1953).
47. Monsalve, R. A., Rogers, A. E. E., Mozdzen, T. J. & Bowman, J. D. One-port direct/reverse method for characterizing VNA calibration standards. *IEEE Trans. Microw. Theory Tech.* **64,** 2631–2639 (2016).
48. Sutinjo, A. T. *et al.* Characterization of a low-frequency radio astronomy prototype array in Western Australia. *IEEE Trans. Antenn. Propag.* **63,** 5433–5442 (2015).
49. Haslam, C. G. T., Salter, C. J., Stoffel, H. & Wilson, W. E. A 408 MHz all-sky continuum survey. II. The atlas of contour maps. *Astron. Astrophys. Suppl. Ser.* **47,** 1–143 (1982).
50. Chandrasekhar, S. *Radiative Transfer* (Courier Dover, 1960).
51. de Oliveira-Costa, A. *et al.* A model of diffuse Galactic radio emission from 10 MHz to 100 GHz. *Mon. Not. R. Astron. Soc*. **388,** 247–260 (2008).
52. Bernardi, G., McQuinn, M. & Greenhill, L. J. Foreground model and antenna calibration errors in the measurement of the sky-averaged λ21 cm signal at $z \sim 20$. *Astrophys. J.* **799,** 90 (2015).
53. Mozdzen, T. J., Bowman, J. D., Monsalve, R. A. & Rogers, A. E. E. Limits on foreground subtraction from chromatic beam effects in global redshifted 21 cm measurements. *Mon. Not. R. Astron. Soc.* **455,** 3890–3900 (2016).
54. McKinley, B. *et al.* Low-frequency observations of the Moon with the Murchison widefield array. *Astron. J.* **145,** 23 (2013).
55. Seager, S., Sasselov, D. D. & Scott, D. How exactly did the Universe become neutral? *Astrophys. J. Suppl. Ser.* **128,** 407–430 (2000).


broadband Blade dipole
Ground plane
Balun
reflection measurement reference plane at input of receiver
RECEIVER Located under Antenna
Noise Source
Cal SW
DC Power
Out of band noise source
26-dB Attenuator
LNA
post AMP
∑
Antenna/Load-Cal SW
Input SW
Circuitry for Measurement of Antenna Reflection
Control
Thermoelectric Temperature Control
COAXIAL RF CABLES
RECEIVER BACKEND located in HUT
multi-conductor cable with RFI filters
Amplification Filtering
spectrometer
14 bit ADC 400 Ms/s
PCI
PC
Vector Network Analyzer
Switch and thermal control
(a)
(b)
(c)
(d)

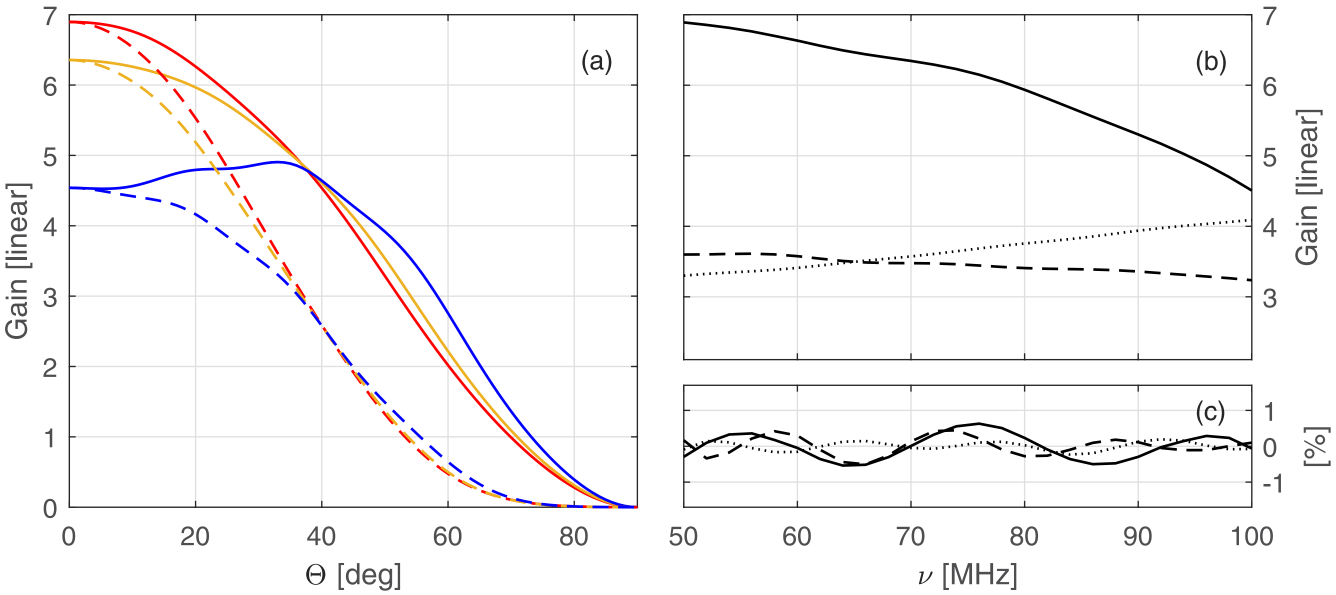
Extended Data Figure 1 | Block diagram of the low-band system. The inset images show: **a**, the capacitive tuning bar that feeds the dipoles at the top of the balun; **b**, the SubMiniature version A connector at the bottom of the balun coaxial transmission line where the receiver connects; **c**, the low-1 receiver that is installed under its antenna with the ground-plane cover plate removed; and **d**, the inside of the low-1 receiver. The LNA is contained in the secondary metal enclosure in the lower-left corner of the receiver. SW, switch.



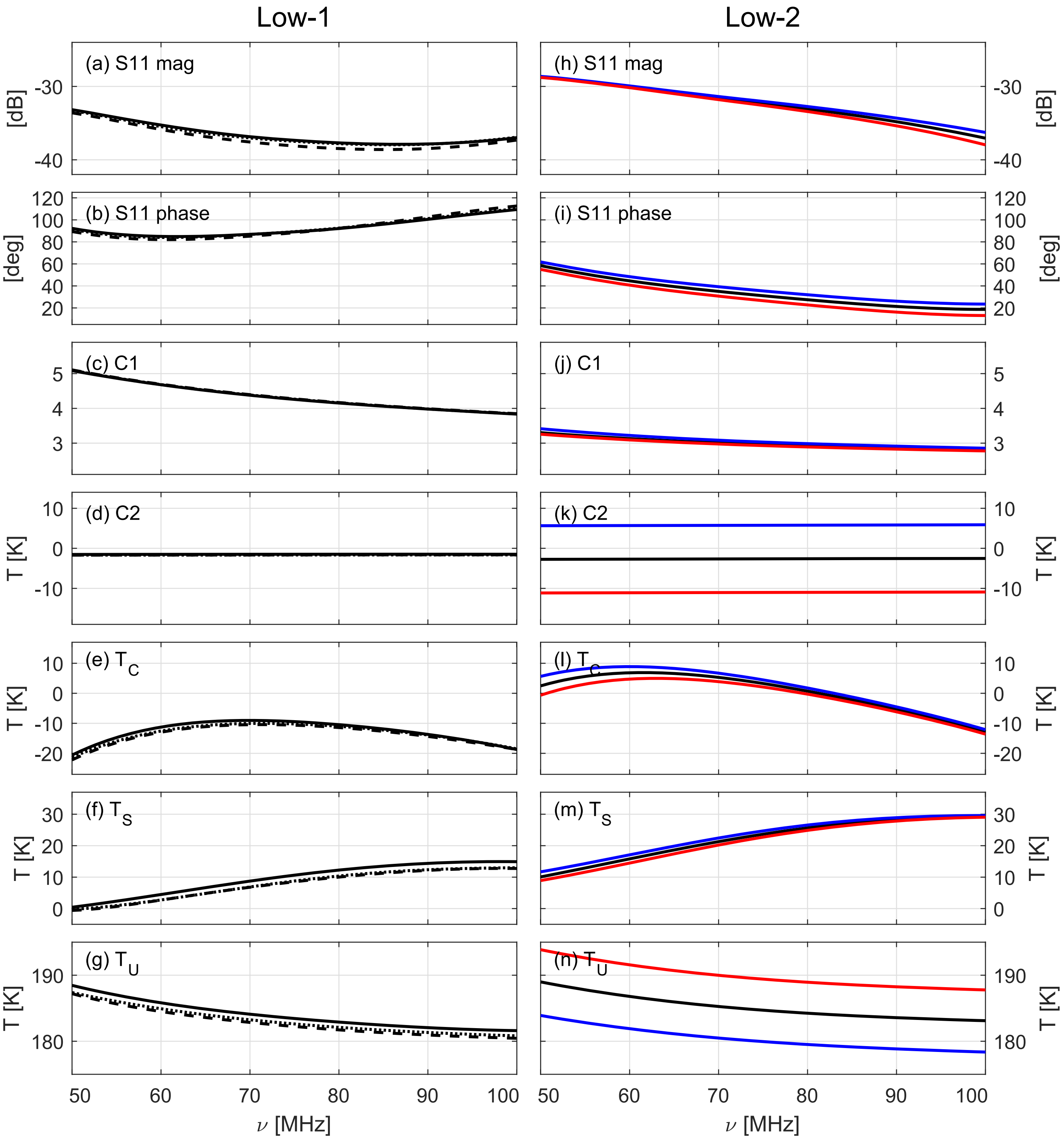
Extended Data Figure 2 | Low-band antennas. a, The low-1 antenna with the 30 m × 30 m mesh ground plane. The darker inner square is the original 10 m × 10 m mesh. The control hut is 50 m from the antenna. **b**, A close view of the low-2 antenna. The two elevated metal panels form the dipole-based antenna and are supported by fibreglass legs. The balun consists of the two vertical brass tubes in the middle of the antenna. The balun shield is the shoebox-sized metal shroud around the bottom of the balun. The receiver is under the white metal platform and is not visible.



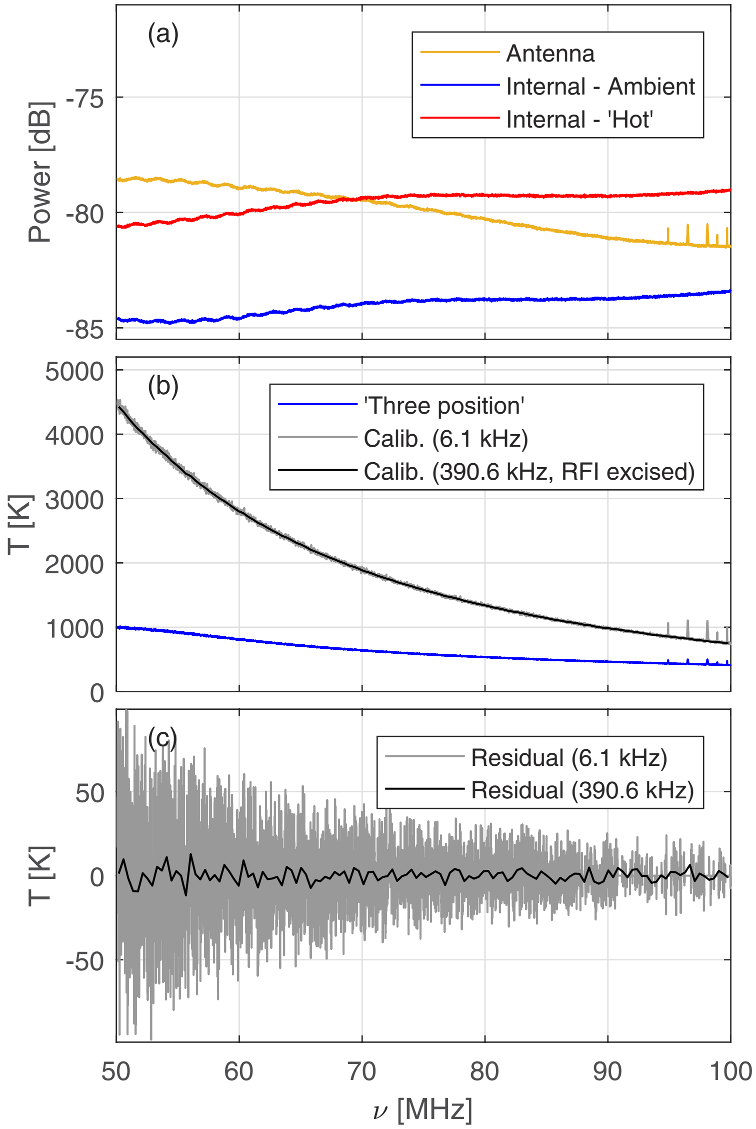
Extended Data Figure 3 | Antenna and simulator reflection coefficients. **a**, **b**, Measurements of the reflection coefficient magnitude (**a**) and phase (**b**) are plotted for hardware configurations H2 (blue), H4 (red) and H6 (yellow). The antennas are designed identically (except H6 has the balun shield removed), but are tuned manually during installation by adjusting the panel separation and the height of the small metal plate that connects one panel to the centre conductor of the balun transmission line on the other. The measurements were acquired *in situ*. **c**, **d**, The red curve is the 10,000-K artificial antenna noise source and the blue curve is the 300-K mismatched load.



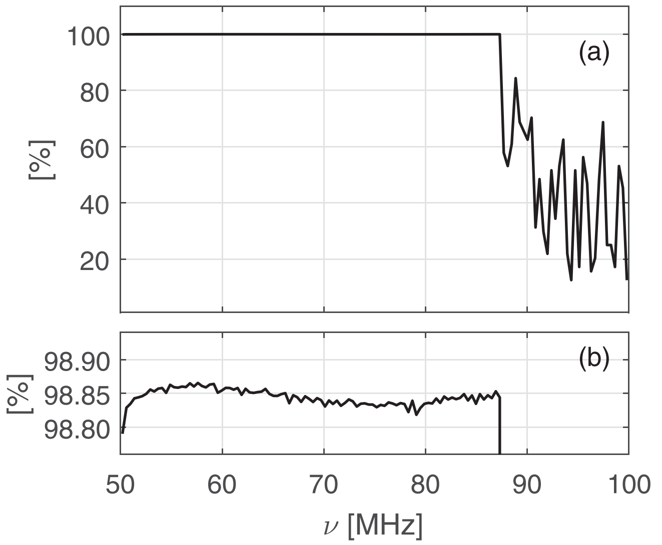
Extended Data Figure 4 | Antenna beam model. a, Beam cross-sections showing the gain in the plane containing the electric field (dashed) and in the plane containing the magnetic field (solid) from FEKO for the H2 antenna and ground plane over soil. Cross-sections are plotted at 50 MHz (red), 70 MHz (yellow) and 100 MHz (blue). **b**, Frequency dependence of the gain at zenith angle $\Theta = 0°$ (solid) and the 3-dB points at 70 MHz in the electric-field plane (dashed) and magnetic-field plane (dotted). **c**, Small undulations with frequency, after a five-term polynomial (equation (2)) has been subtracted from each of the curves, are plotted as fractional changes in the gain. Simulated observations with this model yield residuals of 0.015 K (0.001%) to the five-term fit over the frequency range 52–97 MHz at GHA = 10 and residuals of 0.1 K (0.002%) at GHA = 0, showing that the cumulative beam yields less chromaticity than the approximately 0.5% variations in the individual points plotted.



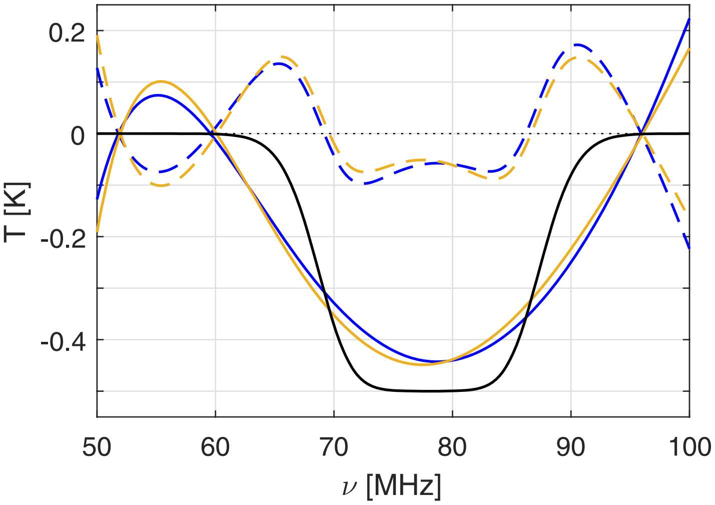
Extended Data Figure 5 | Calibration parameter solutions. **a–g**, Solutions for the low-1 receiver at its fixed 25 °C operating temperature. It was calibrated on three occasions spanning two years, bracketing all of the low-1 observations reported. The first calibration was in August 2015 before commencing cases H1 and H2 (solid), the next was in May 2017 before H3 (dotted), and the final was in September 2017 after the conclusion of H3 (dashed). **h–n**, Solutions for the low-2 receiver controlled to three different temperatures: 15 °C (blue), 25 °C (black) and 35 °C (red). The parameters C1 and C2 are scale and offset factors, respectively; T_C, T_S and T_U are noise-wave parameters (T_S is not the spin temperature here; see ref. 34 for details); S11 is the LNA input reflection coefficient.



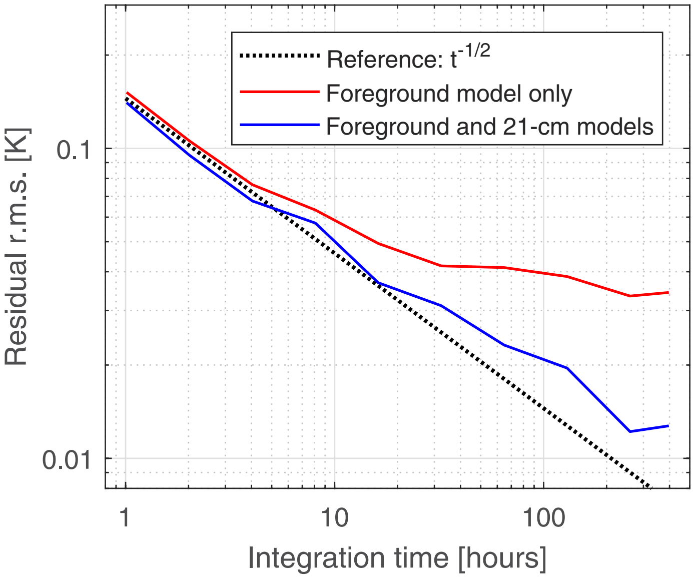
Extended Data Figure 6 | Raw and processed spectra. a, Typical raw 13-s spectra from H2 for each of the receiver's 'three position' switch states. The small spikes on the right of the antenna spectrum are FM radio stations. **b**, The spectrum has been partially calibrated ('Three position') using the three raw spectra to correct gain and offset contributions in the receiver and cables, then fully calibrated (Calib.) by applying the calibration parameter solutions from the laboratory to yield the sky temperature. **c**, Residuals to a fit of the fully calibrated spectrum with the five-term polynomial foreground model (equation (2)). In **b** and **c**, the frequencies listed in the legend give the binning used for each curve.



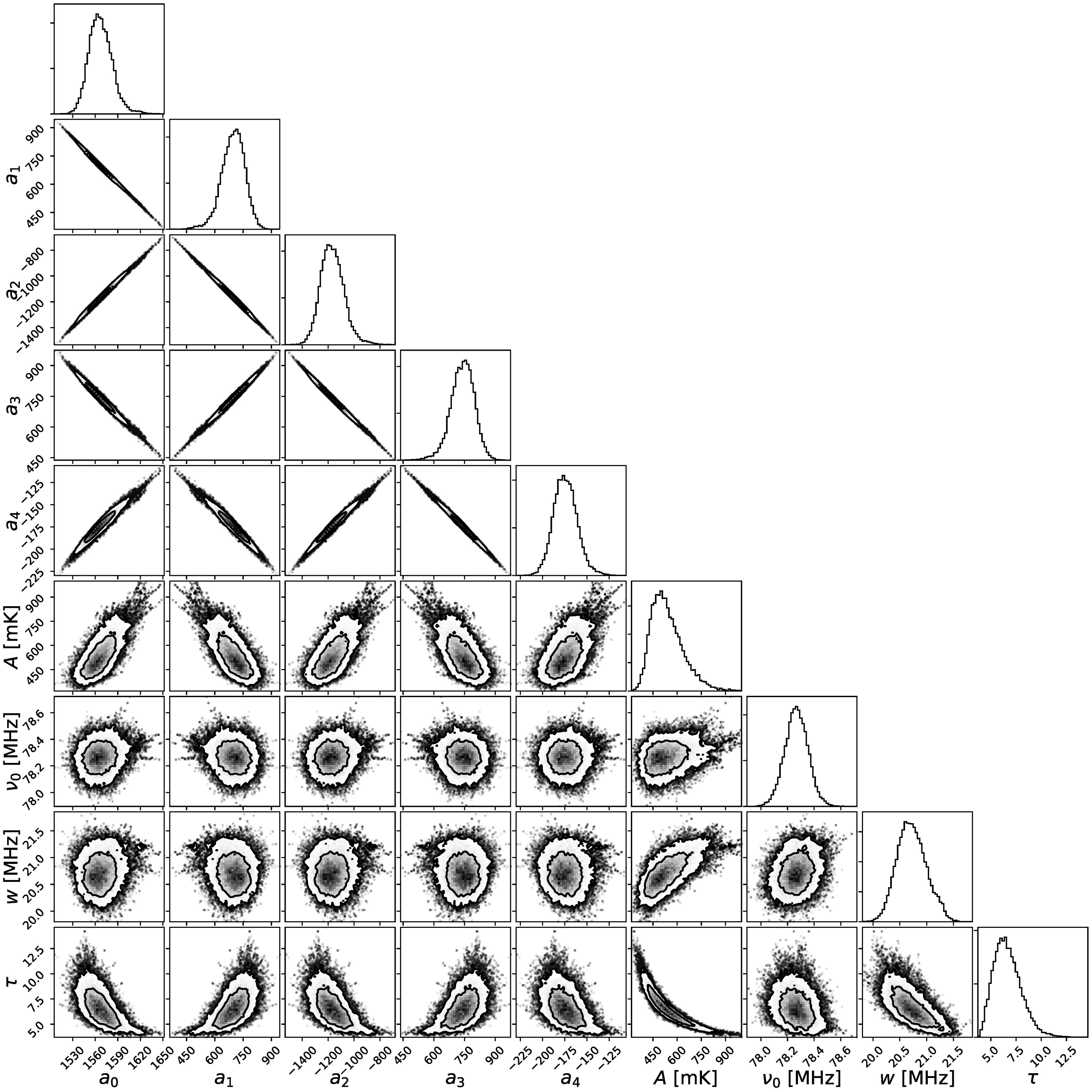
Extended Data Figure 7 | Normalized channel weights. The fraction of data integrated for each 390.6-kHz spectral bin are shown. **a**, The FM band causes the low weights above 87 MHz, because many 6.1-kHz raw spectral channels in this region are removed for all times. The weights are nearly identical across all hardware cases (H1–H6). **b**, A close-up showing the weights below the FM band, where there is little RFI to remove.



Extended Data Figure 8 | Residuals to the 21-cm profile model. The black curve shows the best-fitting 21-cm profile model derived from the observations. The solid blue and yellow curves show fits to the model profile using the physical (equation (1)) and five-term polynomial (equation (2)) foreground models, respectively. The dashed lines show the residuals after subtracting the fits from the model. These residuals are similar to those found when fitting the observations using only a foreground model (Fig. 1b).



Extended Data Figure 9 | Residual r.m.s. as a function of integration time. The curves show the residual r.m.s. after a best-fitting model is removed at each integration time for the H2 dataset.



Extended Data Figure 10 | Parameter estimation. Likelihood distributions for the foreground and 21-cm model parameters are shown for the H2 dataset. Contours are drawn at the 68% and 95% probability levels. The foreground polynomial coefficients (a_n) are highly correlated with each other, whereas the 21-cm model parameters are largely uncorrelated, except for the profile amplitude (A) and flattening (τ). Systematic uncertainties from the verification hardware cases are not presented here.

Extended Data Table 1 | Best-fitting parameter values for the 21-cm absorption profile for representative verification tests

Configuration			Sky Time (hours)	SNR	Centre Frequency (MHz)	Width (MHz)	Amplitude (K)
Hardware configurations (all P6)							
H1 – low-1 10x10 ground plane			528	30	78.1	20.4	0.48
H2 – low-1 30x30 ground plane			428	52	78.1	18.8	0.54
H3 – low-1 30x30 ground plane and recalibrated receiver			64	13	77.4	19.3	0.43
H4 – low-2 NS			228	33	78.5	18.0	0.52
H5 – low-2 EW			68	19	77.4	17.0	0.57
H6 – low-2 EW and no balun shield			27	15	78.1	21.9	0.50
Processing configurations (all H2 except P17)							
P3	–	No beam correction		19	78.5	20.8	0.37
		No beam correction (65-95 MHz)		25	78.5	18.6	0.47
		HFSS beam model		34	78.5	20.8	0.67
		FEKO beam model		48	78.1	18.8	0.50
P4	–	No loss corrections		25	77.4	18.6	0.44
P7	–	5-term foreground polynomial (60-99 MHz)		21	78.1	19.2	0.47
P8	–	Physical foreground model (51-99 MHz)		37	78.1	18.7	0.53
P14	–	Moon above horizon		44	78.1	18.8	0.52
		Moon below horizon		40	78.5	18.7	0.47
P17	–	15°C calibration (61-99 MHz, 5-term)		25	78.5	22.8	0.64
		35°C calibration (61-99 MHz, 5-term)		16	78.9	22.7	0.48

Model fits were performed by using a grid search with fixed $\tau = 7$. Sky time is the amount of time spent by the receiver in the antenna switch state and is 33% of wall-clock time. The data acquisition system has a duty cycle of about 50% and a spectral window function efficiency of about 50%, yielding effective integration times that are a factor of four smaller than the listed sky times. SNR, signal-to-noise ratio.

Extended Data Table 2 | Recovered 21-cm profile amplitudes for various GHAs

Galactic Hour Angle (GHA)	SNR	Amplitude (K)	Sky Temperature (K)
6-hour bins			
0	8	0.48	3999
6	11	0.57	2035
12	23	0.50	1521
18	15	0.60	2340
4-hour bins			
0	5	0.45	4108
4	9	0.46	2775
8	13	0.44	1480
12	21	0.57	1497
16	11	0.59	1803
20	9	0.66	3052

Each block is centred on the GHA listed. The 6-h bins used the five-term physical foreground model (equation (1)) fitted simultaneously with the 21-cm profile amplitude between 64 MHz and 94 MHz. The 4-h bins used a six-term polynomial foreground model (equation (2)) fitted between 65 MHz and 95 MHz. All data are from hardware configuration H2. Sky temperatures are reported at 78 MHz.

doi:10.1038/nature25791

Possible interaction between baryons and dark-matter particles revealed by the first stars

Rennan Barkana[1]

The cosmic radio-frequency spectrum is expected to show a strong absorption signal corresponding to the 21-centimetre-wavelength transition of atomic hydrogen around redshift 20, which arises from Lyman-α radiation from some of the earliest stars[1–4]. By observing this 21-centimetre signal—either its sky-averaged spectrum[5] or maps of its fluctuations, obtained using radio interferometers[6,7]—we can obtain information about cosmic dawn, the era when the first astrophysical sources of light were formed. The recent detection of the global 21-centimetre spectrum[5] reveals a stronger absorption than the maximum predicted by existing models, at a confidence level of 3.8 standard deviations. Here we report that this absorption can be explained by the combination of radiation from the first stars and excess cooling of the cosmic gas induced by its interaction with dark matter[8–10]. Our analysis indicates that the spatial fluctuations of the 21-centimetre signal at cosmic dawn could be an order of magnitude larger than previously expected and that the dark-matter particle is no heavier than several proton masses, well below the commonly predicted mass of weakly interacting massive particles. Our analysis also confirms that dark matter is highly non-relativistic and at least moderately cold, and primordial velocities predicted by models of warm dark matter are potentially detectable. These results indicate that 21-centimetre cosmology can be used as a dark-matter probe.

An excess 21-cm absorption signal is a clear sign of scattering of baryons and dark-matter particles. In general, the intensity of the 21-cm signal is expressed as the observed brightness temperature relative to the cosmic microwave background (CMB), which is given by[1]

$$T_{21} = 26.8 x_{\mathrm{HI}} \frac{\rho_{\mathrm{g}}}{\bar{\rho}_{\mathrm{g}}} \left(\frac{\Omega_{\mathrm{b}} h}{0.0327}\right) \left(\frac{\Omega_{\mathrm{m}}}{0.307}\right)^{-1/2} \left(\frac{1+z}{10}\right)^{1/2} \left(\frac{T_{\mathrm{S}} - T_{\mathrm{CMB}}}{T_{\mathrm{S}}}\right) \qquad (1)$$

in millikelvin. In equation (1), x_{HI} is the mass fraction of neutral (that is, not ionized) hydrogen; ρ_{g} is the gas density and $\bar{\rho}_{\mathrm{g}}$ is its cosmic mean value; Ω_{m} and Ω_{b} are the cosmic mean densities of matter and baryons, respectively, in units of the critical density (the mean density of a flat universe); h is the Hubble parameter in units of 100 km s^{-1} Mpc^{-1}; z is the redshift that corresponds to an observed wavelength of $21(1+z)$ cm and an observed frequency of $1{,}420/(1+z)$ MHz; $T_{\mathrm{CMB}} = 2.725(1+z)$ is the CMB temperature at z; and T_{S} is the spin temperature of hydrogen at z. T_{S} is an effective temperature that describes the relative abundances of the ground and excited states of the hyperfine splitting (spin-flip transition) of the hydrogen atom. In the absence of astrophysical radiation, this temperature is defined by collisions of the hydrogen atoms with each other and scattering of CMB photons[11], and therefore $T_{\mathrm{gas}} \le T_{\mathrm{S}} \le T_{\mathrm{CMB}}$, where T_{gas} is the (kinetic) temperature of the gas.

Observations of the 21-cm line can be used to probe density fluctuations[12], cosmic reionization[13] and X-ray heating[1,14,15], but the earliest observable feature from cosmic dawn is an absorption signal[1–4] that originates from the indirect coupling of T_{S} to T_{gas} by stellar Lyman-α photons via the Wouthuysen–Field effect[16,17]. The first detection of a cosmic 21-cm signal was made by the Experiment to Detect the Global Epoch of reionization Signature (EDGES) [5], which detected the signal's global spectrum from cosmic dawn and found an absorption peak at frequency $\nu = 78 \pm 1$ MHz ($z = 17.2$) with brightness temperature $T_{21} = -500^{+200}_{-500}$ mK; the uncertainties represent 99% confidence intervals and include both thermal and systematic noise. This absorption signal has passed robustness tests for variations in the hardware and processing configuration. If confirmed, this signal (which is 3.8σ below -209 mK, where σ is the standard deviation; the strongest possible absorption at this frequency under standard expectations) cannot be explained without a new dark-matter interaction, even if we take exotic astrophysics into account (see Methods). Indeed, $T_{21} = -300$ mK at $z = 17.2$ implies $T_{\mathrm{gas}} < 5.1$ K, whereas the lowest possible value in the standard scenario is 7.0 K. Basic thermodynamics suggests that it is easy to heat the cosmic gas but difficult to cool it. The extra cooling indicated by the data is possible only through the interaction of the baryons with something even colder.

The only known cosmic constituent that can be colder than the early cosmic gas is dark matter. The reason for this is that dark matter is assumed to interact with itself and with baryons mainly gravitationally, and so it is expected to decouple thermally in the very early Universe and cool down thereafter (very quickly if it is non-relativistic early on, as in the case of cold dark matter). Substantial electrodynamic or nuclear interactions of dark matter would be inconsistent with the observational successes of standard cosmology, including Big Bang nucleosynthesis, CMB observations and the formation and distribution of galaxies. However, weak, non-gravitational interactions are possible. There is a wide array of possibilities for how the strength of such an interaction might vary with temperature or, more specifically, with the velocity v of the scattering baryon relative to the dark-matter particle. Cosmic dawn presents unique physical conditions that can be used to probe a range of parameters that are encountered nowhere else. Specifically, at cosmic dawn, the cosmic gas is at its coldest: it was hotter before owing to its remnant thermal energy from the Big Bang and afterwards owing to X-rays and other heating radiation from astrophysical objects. Therefore, if baryon–dark matter scattering is strongest at low relative velocities, then its effect might be evident only at cosmic dawn.

The cross-section for baryon–dark matter collisions is normally expressed with respect to a relative velocity normalized by the speed of light, and is denoted σ_c. Here, we express it as σ_1, which uses a fiducial relative velocity of 1 km s^{-1}, similar to the typical velocities of baryons and dark-matter particles at cosmic dawn (although in some models they can be less than 0.1 km s^{-1}). We adopt a v^{-4} dependence of the cross-section, which has often been used to illustrate a strongly increasing cross-section with decreasing velocities

$$\sigma(v) = \sigma_c \left(\frac{v}{c}\right)^{-4} = \sigma_1 \left(\frac{v}{1\ \mathrm{km\ s^{-1}}}\right)^{-4} \qquad (2)$$

Such a velocity dependence would arise naturally in the case of Rutherford (or Coulomb) scattering. However, the millicharge model,

[1]Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel.

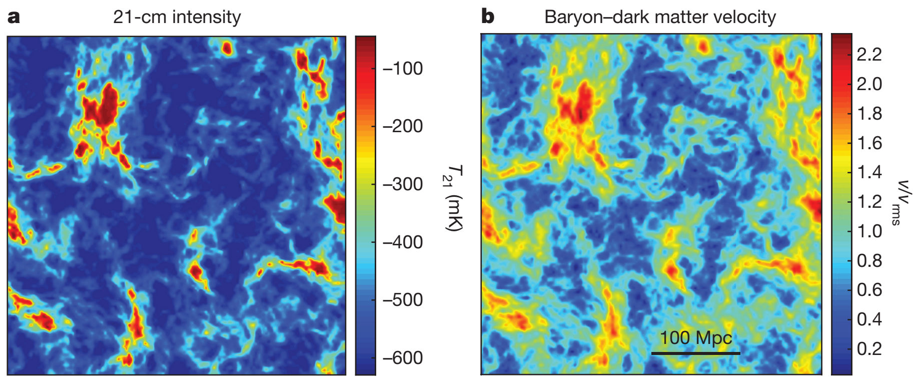


Figure 1 | Simulated 21-cm intensity using a model with baryon–dark matter scattering. a, The 21-cm-transition brightness temperature T_{21} (in millikelvin) in a two-dimensional slice (thickness of 3 Mpc) of a simulated volume of the Universe with sides of 384 Mpc (all lengths co-moving). We consider $z = 17$ ($\nu = 78.9$ MHz), where this model (with $\sigma_1 = 8 \times 10^{-20}$ cm^2 and dark-matter particle mass $m_\chi = 0.3$ GeV) reaches its maximum global 21-cm absorption depth of -504 mK (roughly matching the most likely observed value[5]). This astrophysical model is based on a semi-numerical simulation[15] (see Methods). The spatial 21-cm-signal pattern is determined by the baryon–dark matter relative velocity left over from early cosmic evolution before recombination. **b**, Distribution of the baryon–dark matter relative velocity (in the same simulated volume as that shown in **a**, assuming adiabatic initial density fluctuations), normalized by its r.m.s. value of 29 km s^{-1} at $1 + z = 1{,}010$.

in which dark matter has a small electric charge, is probably ruled out[8,18,19] (see Methods), so we assume a non-standard Coulomb-like interaction between dark-matter particles and baryons that does not depend on whether the baryons are free or bound within atoms.

We calculate the thermal evolution of the baryons and the dark matter by following the exchange of energy and momentum between them[8–10], in which their relative velocity after cosmic recombination has an important role[10]. This velocity remnant[20] arises from the fact that the motion of dark matter is determined by gravity, whereas baryons scatter rapidly off the CMB photons and move along with them in their acoustic oscillations prior to cosmic recombination. This relative velocity (also called the streaming velocity) has received attention recently owing to its effect on early galaxy formation[21], which may produce an observable 21-cm signature[22,23]. However, here we consider baryon–dark matter scattering that depends on the particle velocities directly, not their effect on galaxies.

The baryon–dark matter relative velocity varies spatially (Fig. 1), with a large-scale pattern of coherent regions[21] of about 100 Mpc across. Because the root-mean-square (r.m.s.) velocity is supersonic (decreasing from a Mach number of about 5 right after recombination to approximately 2 when the gas thermally decouples from the CMB) and the scattering cross-section (equation (2)) varies with the relative velocity, the evolution in each region depends strongly on the local value of the initial velocity[10]. At high relative velocities, scattering is weaker (at least until the relative velocity is dissipated by the scattering) and the kinetic energy of the system is partially used to heat the baryons; consequently, a higher relative velocity usually implies less cooling. The dependence of the cooling on velocity results in order-unity fluctuations of the 21-cm intensity(Fig. 1), which we average over (using a Maxwellian distribution for the magnitude of the relative velocity[21]) to estimate the global 21-cm signal.

Such a calculation has been done previously[10], but only for the dark ages, before the formation of any astrophysical sources. In that regime, baryon–dark matter scattering can yield substantial absorption (see the $\nu < 33$ MHz part of the graph in Fig. 2, where previously calculated curves reach a brightness temperature no lower than -70 mK). However, this is unlikely to be observable in the foreseeable future because such low-frequency observations are very difficult to obtain because of ionospheric distortions and because the Galactic synchrotron foreground[3] at $\nu = 20$ MHz is about 40 times stronger than at $\nu = 80$ MHz. A purely cosmological signal would disappear after the dark ages (at $\nu \approx 50$ MHz) because the expansion of the Universe and the cooling of the gas make the coupling of the 21-cm line to $T_{\rm gas}$ (through atomic collisions) less effective than that to the CMB. This drives $T_{\rm S}$ closer to $T_{\rm CMB}$ and eliminates the 21-cm signal.

By combining baryon–dark matter scattering with radiation emitted from the first stars during cosmic dawn in our calculations, we find strong 21-cm absorption that can explain the feature measured by EDGES (Fig. 2). The existence and shape of this absorption dip can be attributed to early astrophysically generated Lyman-α and X-ray radiation backgrounds; this conclusion is consistent with observations that have not detected a strong absorption signal at higher frequencies (see Methods). At the same time, the unexpectedly large depth of the 21-cm absorption indicates cosmic gas that has been cooled substantially by baryon–dark matter scattering. This suggests that only a combination of the two ideas can account for the EDGES data.

The observed 21-cm signal can be explained by considering wide ranges of dark-matter particle masses and baryon–dark matter scattering cross-sections (Fig. 3). Assuming a minimum absorption of -300 mK (as measured[5] at a 99% confidence level), the dark-matter particle must be lighter than 4.3 GeV, which is well below the mass of about 100 GeV that is commonly expected for a weakly interacting massive particle. This finding is consistent with the lack of a direct detection of dark-matter particles by detectors that are sensitive to a wide range of possible weakly interacting massive particles[24,25]. There is no lower limit on the mass except for the extreme limit derived by considering ultra-light ('fuzzy') dark matter[26], namely, $m_\chi \approx 10^{-31}$ GeV, where m_χ is the mass of the dark-matter particle. The same observational result also implies that the scattering cross-section σ_1 is greater than 3.4×10^{-21} cm^2; for the $\sigma(v) \propto v^{-4}$ model (equation (2)), this corresponds to $\sigma_c > 4.2 \times 10^{-43}$ cm^2. Because there is no maximum cross-section, cosmic dawn observations are sensitive to an enormous part of the dark-matter parameter space (in terms of particle mass and cross-section), much of which is unavailable to other probes (see Methods).

The observed signal also places a direct limit on early-Universe scenarios in which dark matter is not completely cold, that is, has a relic thermal velocity. Because dark matter must be colder than baryons in order to cool them, if we demand it to be colder than $T_{[17]}$ (its temperature at $z = 17$), then its r.m.s. velocity at $1 + z = 1{,}010$ must be $v_{\rm rms}^{1010} < (16\,{\rm km\,s^{-1}})\sqrt{T_{[17]}/(5\,{\rm K})}\,\sqrt{(1\,{\rm GeV})/m_\chi}$. In addition, in order for the dark matter to cool the gas at cosmic dawn substantially without disrupting the CMB too much, it must reach a thermal velocity of at most a few kilometres per second at $z = 17$, which puts an upper limit (independent of m_χ) of about 150 km s^{-1} on $v_{\rm rms}^{1010}$. Current limits on warm dark matter[27] allow models with $m_\chi \approx 3$ keV, which have a corresponding $v_{\rm rms}^{1010} \approx 10$ km s^{-1}. Such thermal motion is very far from these upper limits but is comparable to the baryon–dark matter relative velocity, which dominates the 21-cm pattern (Fig. 1); thus, it may be detected or ruled out by observations of 21-cm fluctuations. For

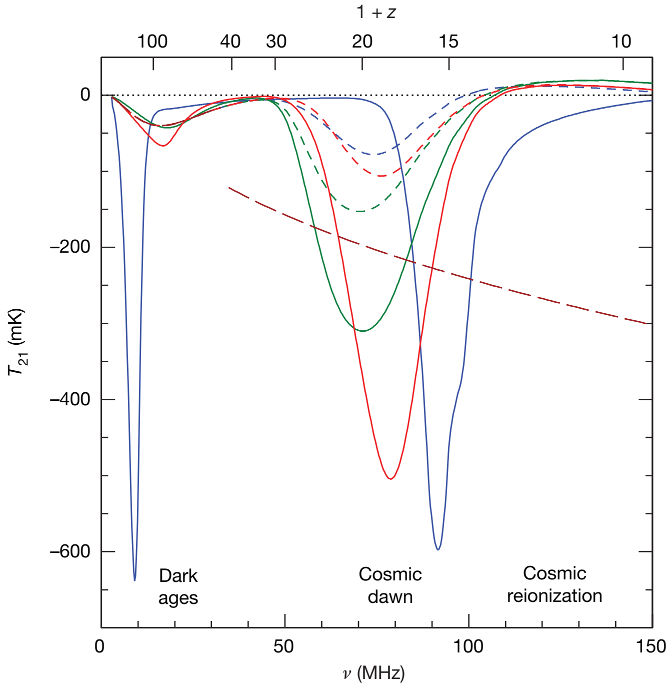


Figure 2 | Global 21-cm signal in models with baryon–dark matter scattering. The globally averaged 21-cm brightness temperature T_{21} (in millikelvin) is shown at an observed frequency ν (in megahertz), with the corresponding value of $1 + z$ displayed at the top. We chart some of the space of possible 21-cm signals (see Methods for a discussion on their shapes) using three models (solid curves), with: $\sigma_1 = 8 \times 10^{-20}$ cm^2 and $m_\chi = 0.3$ GeV (red; roughly matching the most likely observed value[5] of the peak absorption); $\sigma_1 = 3 \times 10^{-19}$ cm^2 and $m_\chi = 2$ GeV (green); and $\sigma_1 = 1 \times 10^{-18}$ cm^2 and $m_\chi = 0.01$ GeV (blue). The astrophysical parameters assumed by these models are given in Methods. The corresponding 21-cm signals in the absence of baryon–dark matter scattering are shown as short-dashed curves. Also shown for comparison (brown long-dashed line) is the standard prediction for future dark ages measurements assuming no baryon–dark matter scattering for $\nu < 33$ MHz (matches all the short-dashed curves in this range) and the lowest global 21-cm signal at each redshift that is possible with no baryon–dark matter scattering, regardless of the astrophysical parameters used (for $\nu > 33$ MHz).

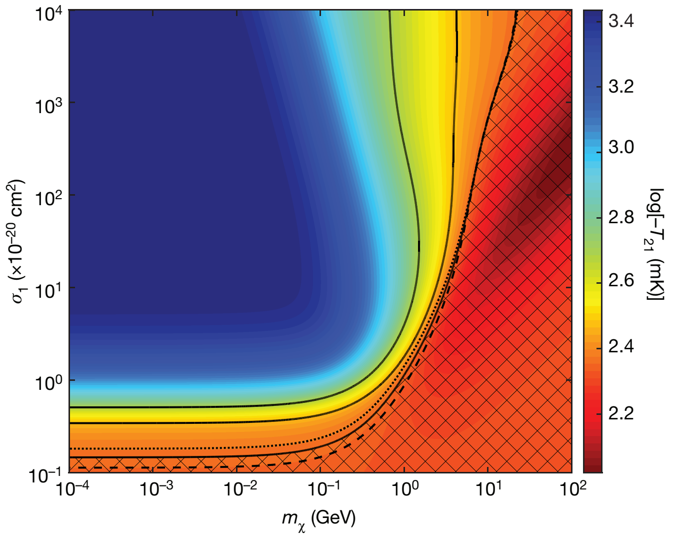


Figure 3 | Constraints on dark-matter properties using cosmic dawn observations. The minimum possible 21-cm brightness temperature T_{21} (expressed as the logarithm of its absolute value) is shown at $z = 17$ ($\nu = 78.9$ MHz), regardless of the astrophysical parameters used (that is, assuming saturated Lyman-α coupling and no X-ray heating), as a function of m_χ and σ_1 (equation (2)). Also shown (solid black curves) are contours corresponding to the following values of T_{21} (from right to left): −231 mK, which corresponds to 10% stronger absorption than the highest value obtained without baryon–dark matter scattering (−210 mK at $z = 17$, or 2.32 on the logarithmic scale); −300 mK, which is the minimal absorption depth in the data at a 99% confidence level; and −500 mK, the most likely absorption depth in the data. The hatched region is excluded if we assume absorption[5] by at least −231 mK at $z = 17$; this 3.5σ observational result implies $\sigma_1 > 1.5 \times 10^{-21}$ cm^2 (corresponding to $\sigma_c > 1.9 \times 10^{-43}$ cm^2 for $\sigma(v) \propto v^{-4}$) and $m_\chi < 23$ GeV. (Although any m_χ above a few gigaelectronvolts requires high σ_1, this parameter combination could be in conflict with other constraints; see Methods.) If we adopt the observed minimum absorption of $T_{21} = -300$ mK, then (again, regardless of astrophysics) the dark matter must satisfy $\sigma_1 > 3.4 \times 10^{-21}$ cm^2 ($\sigma_c > 4.2 \times 10^{-43}$ cm^2) and $m_\chi < 4.3$ GeV; a brightness temperature of −500 mK implies $\sigma_1 > 5.0 \times 10^{-21}$ cm^2 ($\sigma_c > 6.2 \times 10^{-43}$ cm^2) and $m_\chi < 1.5$ GeV. We also illustrate the redshift dependence of these limits via the corresponding 10% contours at $z = 14$ (dashed) and $z = 20$ (dotted).

dark matter that is initially cold, the thermal motion generated by baryon–dark matter scattering may produce effects similar to those predicted by models of warm dark matter (see Methods).

Astronomical testing of the observed signal[5] and of its interpretation in terms of baryon–dark matter scattering will probably begin with other global 21-cm experiments, such as the Shaped Antenna Measurement of the Background Radio Spectrum (SARAS)[28] and the Large-Aperture Experiment to Detect the Dark Ages (LEDA)[29], that will attempt to confirm the measured global signal. Additionally, upcoming 21-cm fluctuation experiments aimed at cosmic dawn will provide a definitive test because the expected spatial pattern of the 21-cm intensity should clearly display a transformed version of the spatial pattern of the baryon–dark matter relative velocity (Fig. 1). Experiments such as the Hydrogen Epoch of Reionization Array (HERA)[6] and the Square Kilometre Array (SKA)[7] should be able to measure the corresponding 21-cm power spectrum because the r.m.s. fluctuation predicted by a model that assumes baryon–dark matter scattering (Fig. 1) is 140 mK (the previously expected maximum value was about 20 mK). Moreover, because of its large spatial scale (of the order of 100 co-moving Mpc, which corresponds to half a degree), the fluctuation pattern should be easy to observe, so no high angular resolution is necessary. As in the case of the galaxy-driven effect of the baryon–dark matter relative velocity[21–23], the power spectrum should show a strong signature of the baryon acoustic oscillations (of order unity in this case) because this velocity arises in part from the participation of baryons in the sound waves of the primordial baryon–photon fluid. A precision measurement at cosmic dawn of the scale of the baryon acoustic oscillations (and thus of the angular diameter distances of the corresponding redshifts) would be a useful cosmological tool to add to current constraints that are based on similar measurements from low-redshift galaxy clustering[30]. If most stars form in galactic haloes with masses lower than about 10^7 solar masses at cosmic dawn, then their spatial distribution should show a similar pattern [21–23] and be strongly anticorrelated with the baryon temperature.

The predicted spatial pattern (Fig. 1) should enable 21-cm imaging of cosmic dawn with the SKA, given the expected sensitivity of the array[7]. The probability distribution function of the 21-cm intensity is expected to be a transformed Maxwellian, which is highly asymmetric, and imaging could verify this unanticipated non-Gaussianity directly. Because the presence of dark matter has historically been inferred from the general theory of relativity on galactic and cosmological scales, confirmation of the existence of dark matter would constitute not only a discovery of physics beyond the standard model, but also verification of this theory.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 July 2017; accepted 23 January 2018.

1. Madau, P., Meiksin, A. & Rees, M. J. 21 centimeter tomography of the intergalactic medium at high redshift. *Astrophys. J.* **475,** 429–444 (1997).
2. Tozzi, P., Madau, P., Meiksin, A. & Rees, M. J. Radio signatures of H i at high redshift: mapping the end of the "dark ages". *Astrophys. J.* **528,** 597–606 (2000).
3. Furlanetto, S. R., Oh, S. P. & Briggs, F. H. Cosmology at low frequencies: the 21 cm transition and the high-redshift Universe. *Phys. Rep.* **433,** 181–301 (2006).
4. Barkana, R. The rise of the first stars: supersonic streaming, radiative feedback, and 21-cm cosmology. *Phys. Rep.* **645,** 1–59 (2016).
5. Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J. & Mahesh, N. An absorption profile centred at 78 megahertz in the sky-averaged spectrum. *Nature* **555,** https://doi.org/10.1038/nature25792 (2018).
6. DeBoer, D. R. *et al.* Hydrogen Epoch of Reionization Array (HERA). *Publ. Astron. Soc. Pacif.* **129,** 045001 (2017).
7. Koopmans, L. *et al.* The cosmic dawn and epoch of reionisation with SKA. In *Proc. Advancing Astrophysics with the Square Kilometre Array* (Proceedings of Science, 2015).
8. Dvorkin, C., Blum, K. & Kamionkowski, M. Constraining dark matter–baryon scattering with linear cosmology. *Phys. Rev. D* **89,** 023519 (2014).
9. Tashiro, H., Kadota, K. & Silk, J. Effects of dark matter-baryon scattering on redshifted 21 cm signals. *Phys. Rev. D* **90,** 083522 (2014).
10. Muñoz, J. B., Kovetz, E. D. & Ali-Haïmoud, Y. Heating of baryons due to scattering with dark matter during the dark ages. *Phys. Rev. D* **92,** 083528 (2015).
11. Purcell, E. M. & Field, G. B. Influence of collisions upon population of hyperfine states in hydrogen. *Astrophys. J.* **124,** 542–549 (1956).
12. Hogan, C. J. & Rees, M. J. Spectral appearance of non-uniform gas at high *z*. *Mon. Not. R. Astron. Soc.* **188,** 791–798 (1979).
13. Scott, D. & Rees, M. J. The 21-cm line at high redshift: a diagnostic for the origin of large scale structure. *Mon. Not. R. Astron. Soc.* **247,** 510–516 (1990).
14. Furlanetto, S. R. The global 21-centimeter background from high redshifts. *Mon. Not. R. Astron. Soc.* **371,** 867–878 (2006).
15. Fialkov, A., Barkana, R. & Visbal, E. The observable signature of late heating of the Universe during cosmic reionization. *Nature* **506,** 197–199 (2014).
16. Wouthuysen, S. A. On the excitation mechanism of the 21-cm (radio-frequency) interstellar hydrogen emission line. *Astron. J.* **57,** 31–32 (1952).
17. Field, G. B. Excitation of the hydrogen 21-cm line. *Proc. IRE* **46,** 240–250 (1958).
18. Chuzhoy, L., & Kolb, E. W. Reopening the window on charged dark matter. *J. Cosmol. Astropart. Phys.* **7,** 14 (2009).
19. McDermott, S. D., Yu, H.-B. & Zurek, K. M. Turning off the lights: how dark is dark matter? *Phys. Rev. D* **83,** 063509 (2011).
20. Sunyaev, R. A. & Zeldovich, Y. B. Small-scale fluctuations of relic radiation. *Astrophys. Space Sci.* **7,** 3–19 (1970).
21. Tseliakhovich, D. & Hirata, C. Relative velocity of dark matter and baryonic fluids and the formation of the first structures. *Phys. Rev. D* **82,** 083520 (2010).
22. Dalal, N., Pen, U.-L. & Seljak, U. Large-scale BAO signatures of the smallest galaxies. *J. Cosmol. Astropart. Phys.* **11,** 7 (2010).
23. Visbal, E., Barkana, R., Fialkov, A., Tseliakhovich, D. & Hirata, C. M. The signature of the first stars in atomic hydrogen at redshift 20. *Nature* **487,** 70–73 (2012).
24. XENON Collaboration. First dark matter search results from the XENON1T experiment. *Phys. Rev. Lett.* **119,** 181301 (2017).
25. PandaX-II Collaboration. Dark matter results from 54-ton-day exposure of PandaX-II experiment. *Phys. Rev. Lett.* **119,** 181302 (2017).
26. Hu, W., Barkana, R. & Gruzinov, A. Fuzzy cold dark matter: the wave properties of ultralight particles. *Phys. Rev. Lett.* **85,** 1158–1161 (2000).
27. Iršič, V. *et al.* New constraints on the free-streaming of warm dark matter from intermediate and small scale Lyman-α forest data. *Phys. Rev. D* **96,** 023522 (2017).
28. Singh, S. *et al.* First results on the epoch of reionization from first light with SARAS 2. *Astrophys. J.* **845,** L12 (2017).
29. Bernardi, G., McQuinn, M. & Greenhill, L. J. Foreground model and antenna calibration errors in the measurement of the sky-averaged λ21 cm signal at $z \sim 20$. *Astrophys. J.* **799,** 90 (2015).
30. Alam, S. *et al.* The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *Mon. Not. R. Astron. Soc.* **470,** 2617–2652 (2017).

Acknowledgements I am grateful to J. Bowman for alerting me to possible indications of very deep absorption in the EDGES results, which inspired this work. This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to R.B. (barkana@tau.ac.il).

METHODS

The measured signal and its theoretical interpretation. In this work we relied on the EDGES measurement of the global 21-cm spectrum from cosmic dawn[5]. The absorption signal was detected with a signal-to-noise ratio of 37. Moreover, the signal was observed in data spanning nearly two years and has passed many robustness tests with little change[5]. Various hardware configurations were tried, including two separately built copies of the instrument, various sizes of the metal ground plane (from 10 m to 30 m on a side), variations in the instrument's orientation and inclusion (or not) of a balun shield. Also, the processing configurations included: two independent processing pipelines (tested with simulated data); various calibration techniques and measurements; inclusion (or not) of beam and balun or ground-plane loss corrections; a four-, five- or six-term foreground polynomial or a different physically motivated five-term foreground model; various frequency intervals (51–99 MHz, 61–99 MHz or 65–95 MHz); and data collected at various sky positions of the Sun, Moon and Galaxy. The measured absorption profile is inconsistent with typical spectra of radio-frequency interference, does not appear Galactic (it is inconsistent with the absorption spectra of H II regions or radio-frequency recombination lines and is not concentrated in the Galactic plane), and is not explained by the ionosphere (which produces a broadband absorption with diurnal variations, as shown by models and observations) or by molecular lines in the atmosphere (which are much too weak).

For the theoretical predictions, we combined a calculation of the effect of baryon–dark matter scattering with a simulation of the effect of the first stars on intergalactic hydrogen. We calculated the thermal exchange between baryons and dark matter by following the evolution of the baryon–dark matter relative velocity (equation (20) in ref. 10) and including heating and cooling terms related to scattering (equations (18) and (19) of ref. 10, using also equations (13), (14) and (16) of the same paper). We treated the baryons as equal-mass particles with 1.22 times the proton mass (which is the mean molecular mass of neutral primordial gas), whereas in ref. 10 one proton mass is used. Thus, equation (2) effectively represents the cross-section of the scattering between dark matter and an average baryon. In reality, the treatment of helium is probably complicated and highly model-dependent[8]. We started our calculation at kinematic decoupling ($1+z=1{,}010$), as in previous calculations[10], and we confirmed that starting earlier would not affect our results much at lower redshifts.

In addition to baryon–dark matter scattering, at each redshift we included spatially uniform backgrounds of astrophysical radiation of the three types that are important in 21-cm cosmology (Lyman-α photons, X-rays and ionizing photons). More precisely, we used the volume-averaged values of Lyman-α coupling (before low-temperature corrections), the rate of heating due to X-rays and the ionized fraction, all taken from a semi-numerical simulation[15,31] with astrophysical parameters chosen to illustrate absorption dips that are consistent with the observed signal. In the model used to obtain the results shown in Fig. 1 and the red curve in Fig. 2, star formation occurs only in haloes that allow atomic cooling and with an efficiency of $f_*=1.58\%$, and X-rays normalized on the basis of low-redshift observations are emitted with a soft power-law spectrum. The green curve in Fig. 2 is obtained with the same model, except that the efficiency of the production of Lyman-α photons is 10 times higher. The model that gives the blue curve in Fig. 2 assumes the same parameters as the model that obtains the red curve but considers an efficiency of $f_*=0.5\%$ and assumes that star formation occurs in haloes that allow molecular cooling. Moreover, it considers an X-ray efficiency 4 times higher than that of the model of the red curve and an X-ray spectrum that extends down to 0.1 keV instead of 0.2 keV. Astrophysical radiation fields are expected to vary spatially, leading to 21-cm fluctuations during cosmic dawn due to Lyman-α fluctuations[32] and X-ray heating fluctuations[33]; these are sizeable and potentially observable. However, we neglected these fluctuations here because of the much larger ones that result from baryon–dark matter scattering. Throughout the analysis, we assumed the known values of the cosmological parameters[34], $\Omega_\mathrm{m}=0.307$, $\Omega_\mathrm{b}=0.0482$ and $h=0.678$.

In most of the model space, the theoretically predicted dip in the cosmic-dawn absorption curve shown in Fig. 2 is well fitted by a simple Gaussian (although for the blue solid curve in Fig. 2, this is true only for the deepest portion of the absorption feature). The measurement of ref. 5 favours a different flattened-Gaussian shape but systematic variations in the hardware configuration weaken this conclusion. In figure 2 of ref. 5, two of the six plotted best-fit profiles and residuals are not very flattened. The data fits show an anticorrelation between the degree of flattening and the absorption amplitude; low flattening (as suggested by the theoretical models) favours high amplitudes of around 1,000 mK.

Strongest possible absorption without baryon–dark matter scattering. A measurement of stronger-than-expected absorption is evidence of the existence of dark matter, as such absorption cannot be produced without baryon–dark matter scattering. In the standard picture, the best-case scenario for producing strong 21-cm absorption is to assume no reionization ($x_{\mathrm{H\,I}}=1$ in equation (1)), saturated coupling ($T_\mathrm{S}=T_\mathrm{gas}$) and no astrophysical heating. In this case, the gas at high redshifts is colder than the CMB because its adiabatic cooling is faster. However, the baryons are thermally coupled to the CMB through Compton heating until $z\approx150$. This well understood phenomenon yields[35] a strongest possible absorption signal (regardless of the uncertain astrophysics at high redshift) of $T_{21}=-209$ mK at $\nu=78$ MHz. We note that this maximum possible absorption is an extreme value (in the standard case without baryon–dark matter scattering) that would not be considered very likely. Models with various astrophysical parameter values[31] predict T_{21} values at $\nu=78$ MHz that range from -209 mK up to positive values, with most of them between -150 mK and -50 mK. More generally, the most negative global 21-cm signal at each frequency that would be possible with no baryon–dark matter scattering (regardless of the parameters of high-redshift astrophysics) is shown (at $\nu>33$ MHz) by the brown long-dashed curve in Fig. 2.

We consider various ideas for increasing the absorption without baryon–dark matter interactions. Fluctuations in the gas density ρ_g affect the 21-cm signal, as the absorption strength is proportional to ρ_g. However, adiabatic heating with $T_\mathrm{gas}\propto\rho_\mathrm{g}^{2/3}$ counteracts this and leads to only a small increase in the absorption in overdense regions, whereas in the voids these factors combine to weaken the overall absorption. Actually, linear fluctuations are symmetric and cancel out when averaged globally over overdense and underdense regions. To change the observed global signal, nonlinear fluctuations are needed. The regime of mildly nonlinear density fluctuations is well understood, as it corresponds to the sheets and filaments of the cosmic web that successfully explain[36] the observed properties of the Lyman-α forest at $z=2$–5. At cosmic dawn, the Universe was probably much more homogeneous than at $z=2$–5 and had less-nonlinear density fluctuations because gravity had not had as much time to drive the growth of fluctuations. Nevertheless, even if we were to assume that somehow the density fluctuations corresponding to the Lyman-α forest were already in place at $z\approx20$, this still would not produce a larger absorption signal than in the absence of such density fluctuations. To check this quantitatively, we assume the best case of $T_\mathrm{S}=T_\mathrm{gas}$ and adiabatic heating and cooling, and we average the 21-cm brightness temperature over the density distribution at $z=2$–6 in simulations that match Lyman-α observations[37]. The result is a weaker average absorption than would occur in the absence of density fluctuations. More evidence that density fluctuations do not produce unusual absorption comes from numerical simulations of the Universe at cosmic dawn; these have been run on various volumes and at various resolutions[38–40], and none of them has predicted a stronger globally averaged absorption signal than the simple limit shown by the brown long-dashed curve in Fig. 2.

Under standard cosmology, the total gas fraction within virialized haloes at $z=20$ is expected to be below 1% because the intergalactic gas can only collect in haloes of mass at least equal to the filtering mass[41,42] of about 3×10^4 solar masses. We can consider, though, an exotic scenario where unexpectedly large density fluctuations on small scales would produce a much larger abundance of early haloes. This would not produce more absorption either. The lowest T_gas at $z=20$ in the standard scenario is about 9 K at the cosmic mean density. As the gas is heated adiabatically, it reaches the CMB temperature (57 K at $z=20$) at an overdensity of 16. After that point it contributes extra emission, not absorption. When the gas enters a virialized halo, it is probably shock-heated. If it cools efficiently, primordial cooling via molecular hydrogen is effective only down to temperatures of a few hundred kelvin and, in any case, efficient cooling probably leads to star formation and even more heating. We also note that the 21-cm optical depth of the coldest-possible gas (without baryon–dark matter scattering) is $\tau_{21}\approx10\%$ at the mean density at $z=20$; this varies as $\tau_{21}\propto\rho_\mathrm{g}/T_\mathrm{gas}\propto\rho_\mathrm{g}^{1/3}$, assuming adiabatic evolution. This means that only very dense gas (inside virialized haloes) can be optically thick, and such gas is expected to be hot.

Another possibility is to change the residual electron fraction after recombination, which determines the rate of Compton heating that keeps the gas close to the CMB temperature until $z\approx150$. To produce unusually strong absorption, such as $T_{21}=-300$ mK at $\nu=77$ MHz, the gas would need to decouple thermally at a value of $1+z$ that is larger by a factor of 1.4 than that predicted by standard cosmology; this would happen if the residual ionized fraction were lower than expected by a factor of about 4. Before cosmic recombination, the gas is strongly coupled to the CMB and cannot cluster, so it would probably be unaffected even by exotic physics, such as unusually strong dark-matter clumping. After the freeze-out at the end of cosmic recombination, the recombination time continues to increase as $1/\rho_\mathrm{g}$, so the residual electron fraction changes slowly with time and is only weakly dependent on density (in part because the recombination coefficient declines with temperature and the temperature rises with density). It is difficult to imagine something that could lower the mean residual electron fraction by a large factor.

More generally, it would be difficult to change the basic cosmological parameters, the cosmic expansion history or the physics involved in cosmic recombination substantially. These inputs are strongly constrained by the success of standard cosmology in fitting observations of the CMB plus low-redshift observations. Proposed exotic astrophysics or physics concepts—such as unexpected populations of stars or black holes, or dark-matter annihilation or decay—also cannot explain the stronger absorption. Such scenarios would generate extra ultraviolet, X-ray or γ-ray radiation, which would generate more heating as well as more ionization (which would lower $x_{\rm HI}$ and also boost the Compton heating of the gas). Also, Lyman-α coupling cannot be stronger than the saturated coupling limit ($T_{\rm S} = T_{\rm gas}$) that we have considered here.

Astrophysical considerations and implications. While the detailed parameters of the astrophysical sources at high redshift are highly uncertain, strong 21-cm absorption is a generic prediction. A scan through a wide range of currently plausible astrophysical parameter values[31] (without baryon–dark matter scattering) shows that all models feature an absorption dip at cosmic dawn[14], which is produced (in the direction of increasing ν) by a fall (that is, increasing absorption) due to increasing Lyman-α coupling, followed by a rise caused by increasing X-ray heating (or due to reionization in models with late X-ray heating). The depth of the absorption dip[31] is in the range $-240\,{\rm mK} < T_{21,{\rm min}} < -25\,{\rm mK}$, and its position is in the range $52\,{\rm MHz} < \nu_{\rm min} < 120\,{\rm MHz}$ (corresponding to $11 < z_{\rm min} < 26.5$).

Once baryon–dark matter scattering is included, the observed global 21-cm signal is determined by the complex interplay of this scattering with astrophysics. For example, in the large region (Fig. 3) of low m_χ and high σ_1 that is allowed, the initial cooling due to baryon–dark matter scattering can be extremely effective and lead to global 21-cm absorption as strong as -600 mK in the dark ages (but only at very high redshifts, higher than 100). In these models, the gas is so cold that Lyman-α coupling is delayed owing to low-temperature corrections (discussed below), and cosmic heating is also delayed because X-ray heating must initially counteract the baryon–dark matter cooling. In this region of parameter space, astrophysical models can be chosen to produce an absorption peak position and depth that agree with the data[5], but detailed parameter constraints require full consideration of the large variety of possible astrophysical parameters, which we leave for future work. We also note that a very high σ_1 would tend to suppress the baryon–dark matter relative velocity and, with it, the associated fluctuations (discussed in the main text); although the normal 21-cm fluctuations due to inhomogeneous galaxy formation would be enhanced in proportion to the (unexpectedly large) absolute value of the mean global signal.

The observed global 21-cm signal[5] represents the first detection of some of the earliest stars. The location of the peak absorption at $z \approx 17$ is not surprising, but it considerably narrows down astrophysical parameters that were previously almost unconstrained. In general, the maximal absorption corresponds to the late stages of Lyman-α coupling, together with the early stages of X-ray heating. The observed timing of these astrophysical cosmic milestones is well within the expected range of astrophysical parameters[31], which further supports the dark-matter-cooling interpretation, as there is no indication of exotic astrophysics. Interestingly, this early heating is also consistent with limits (obtained from 21-cm observations during the epoch of reionization) on both the global and fluctuation signals[28,43,44], which disfavour strong absorption at low redshifts (such absorption is expected in the case of late heating). Indeed, the detected signal implies that future 21-cm observations should focus on cosmic dawn, where the 21-cm signal (both the global signal and the power spectrum) is probably much stronger than previously expected, and not on the later era of cosmic reionization, which has been the focus thus far[28,43–47] and where the signal strength is probably in the lower part of the previously expected range.

The Lyman-α coupling of the 21-cm line is known to become less effective when the gas temperature is low[4,48–51]. In the previously standard case, these low-temperature corrections amount at most to a 20% reduction in the coupling at any redshift[4]. With the lower gas temperatures encountered in the case of substantial baryon–dark matter scattering, however, the low-temperature corrections can reduce the coupling by an order of magnitude or more, delaying strong Lyman-α coupling and greatly changing the global 21-cm signal. Indeed, in some models the gas temperature is so low (<0.1 K) that these low-temperature corrections (as well as additional basic physics of 21-cm cosmology) may need to be re-assessed. Furthermore, the standard expression for the 21-cm signal (equation (1)) is a linearization that assumes a low 21-cm optical depth (a valid assumption if there is no baryon–dark matter scattering), but here we encounter high optical depth values and thus use the more general expression[1]. Lyman-α scattering can also affect the thermal state of the cosmic gas; for almost all reasonable astrophysical parameters[49,52], heating due to continuum photons dominates over cooling from injected photons (for gas below about 100 K) and the heating rate is weak compared to that of X-ray heating.

Substantial baryon–dark matter scattering would also influence the formation of the first stars. The lower gas temperature would reduce the Jeans mass, and the dissipation of the baryon–dark matter relative velocity would reduce its suppression effects on star formation. Both of these effects would boost star formation relative to the case of no baryon–dark matter scattering, but their impact might be limited because of the large time scales needed for galaxy formation. Baryon infall into dark-matter haloes begins at recombination, and for most dark-matter parameter values it takes some time until the baryon–dark matter scattering has a considerable effect. We have neglected effects on galaxy formation in this work because they are dwarfed by the direct effect of excess gas cooling on the 21-cm signal.

The range of dark-matter properties that can be probed by cosmic dawn observations. The dark-matter parameters that affect 21-cm cosmology are shown in Fig. 3, but it is important to understand the basic physics behind these constraints. In particular, we can understand why there is an upper limit on the mass of a dark-matter particle that can cool cosmic baryons by considering the maximum possible cooling. As mentioned previously, baryons thermally decoupled from the CMB at $z \approx 150$. In the presence of baryon–dark matter scattering, by that time the dark matter had acquired a non-zero temperature T_χ; however, $T_\chi \ll T_{\rm gas}$ is required for maximum cooling of the baryons at this time. For cooling to occur, the two fluids must be strongly coupled after the baryons thermally decouple, so that the baryons share some of their energy with the dark matter. The most that such coupling can achieve, if it is strong, is a thermal equilibrium, at which both the baryons and the dark matter have a common (time-dependent) temperature $T_{\rm fin}$. Then at a given time, the relation between the baryon temperature $T_{\rm gas}$ in the absence of baryon–dark matter scattering and the lowest possible temperature $T_{\rm fin}$ with scattering is given by conservation of energy (per unit volume) as

$$T_{\rm fin} = T_{\rm gas}\frac{n_{\rm b}}{n_{\rm b} + n_\chi} = \frac{T_{\rm gas}}{1 + (\rho_\chi/\rho_{\rm b})(\mu_{\rm b}/m_\chi)} \approx \frac{T_{\rm gas}}{1 + (6\ {\rm GeV})/m_\chi} \qquad (3)$$

where $n_{\rm b}$ and n_χ are the number densities of baryons and dark matter, respectively, $\rho_{\rm b}$ and ρ_χ are the corresponding (mean) densities, $\mu_{\rm b}$ is the mean baryonic mass and m_χ is the mass of a dark-matter particle. Although we have neglected the effect of the initial baryon–dark matter relative velocity, the kinetic energy associated with it would only produce more heating. As an example, to reach $T_{21} = -300$ mK at $z = 17$ (the current limit at 99% confidence[5]), the simple estimate in equation (3) yields a maximum possible m_χ of 16 GeV. In reality, the cooling never reaches the best-case scenario assumed in this simple estimate, and we find an actual maximum mass of 4.3 GeV (see Fig. 3).

There is no lower limit on the mass of the dark-matter particle that can affect the 21-cm signal, because the cooling rate becomes independent of m_χ when $m_\chi \ll \mu_{\rm b}$. In that limit, the energy lost by a baryon per collision (at a given baryon–dark matter relative velocity) is proportional to m_χ and the scattering rate is proportional to $n_\chi\sigma_1$; therefore, the total cooling rate is proportional to $\rho_\chi\sigma_1$, where ρ_χ is the known mean density of dark matter. Thus, a substantial interaction requires a minimum σ_1 that is independent of m_χ when $m_\chi \ll \mu_{\rm b}$ (Fig. 3). We note that the enormous range of particle masses (as well as scattering cross-sections) that are potentially detectable by cosmic dawn observations can be considered an argument that such a detection is not an unreasonable possibility.

The dependence of the effectiveness of the baryonic cooling on the baryon–dark matter scattering cross-section is non-trivial. A higher cross-section means that more of the thermal energy of the baryons can be transferred to the dark matter; on the other hand, it also implies that the dark matter warms up earlier on, before thermal decoupling of the gas from the CMB, which reduces the ability of the dark matter to later cool the gas. There is even a region (for example, $\sigma_1 = 2 \times 10^{-18}\ {\rm cm}^2$ and $m_\chi = 100$ GeV; see Fig. 3) where the baryon–dark matter interaction causes a small net baryonic heating due to another effect, namely, the transfer of kinetic energy from their relative velocity to the random gas motions. In the limit of very low m_χ and very high σ_1 (the top-left portion of Fig. 3), the absorption (which is defined with respect to the CMB at zero redshift) approaches its maximum possible value, namely, the CMB temperature of 2.725 K.

Another issue is the effect of baryon–dark matter scattering on the dark matter. According to equation (3), the maximum velocity dispersion, v_χ, of the dark matter (even if it is initially completely cold) relative to the velocity dispersion of the gas, $v_{\rm b}$, in the absence of baryon–dark matter scattering is:

$$\frac{v_\chi}{v_{\rm b}} = \sqrt{\frac{T_{\rm fin}/m_\chi}{T_{\rm gas}/\mu_{\rm b}}} = \left(5 + \frac{m_\chi}{1.1\ {\rm GeV}}\right)^{-1/2} \qquad (4)$$

In reality, the dark matter can have a velocity dispersion that is very different at a given time from that corresponding to $T_{\rm fin}$ of equation (3). A higher v_χ can be caused by memory of earlier times—in particular, of the higher temperature of the gas when it was thermally coupled to the CMB—because after decoupling the dark matter may not be able to cool fast enough to forget its history effectively (the fastest it can cool is adiabatically). Later, once X-rays (and eventually reionization) heat the gas, v_χ can remain low even when v_b rises, as the increasing v_b weakens the baryon–dark matter scattering (assuming the $\sigma(v) \propto v^{-4}$ model). Numerically, we find that v_χ decreases with time but the matter power spectrum may be suppressed by the thermal velocity that exists at high redshifts, when linear growth critical for the later emergence of structure occurs. This may offer a way to alleviate the small-scale crisis of cold dark matter (that is, the model's difficulties in matching the observed properties of dwarf galaxies) that has led to models such as those of warm dark matter and fuzzy dark matter[26,27,53], and can help to set upper limits on the baryon–dark matter cross-section[8]. Allowed models of warm dark matter have $v_{\mathrm{rms}} \approx 5\,\mathrm{km\,s^{-1}}$ at $z = 500$, so we estimate that models with a similar v_χ at that redshift may also have a considerable effect on the Lyman-α forest or on dwarf galaxies at low redshift. In regions with zero initial streaming velocity, this value of v_χ occurs for $\sigma_1 \approx 10^{-18}\,\mathrm{cm}^2$ as long as $m_\chi \ll 1$ GeV (the σ_1 required rises by a factor of 10 for $m_\chi = 1$ GeV); if σ_1 is ten times lower, then v_χ drops to $2\,\mathrm{km\,s^{-1}}$ at $z = 500$.

We have assumed throughout this work that $\sigma(v) \propto v^{-4}$ but the velocity dependence of the cross-section can be explored with future 21-cm data. Further global 21-cm measurements may help but will probably not resolve the degeneracies between the dark-matter parameters (cross-section amplitude, velocity dependence and particle mass) and the astrophysical parameters (the X-ray spectrum and the normalization and redshift evolution of the various radiation backgrounds, as determined by parameters related to galaxy formation). Detailed measurements of 21-cm fluctuations, including the 21-cm power spectrum, would provide much more information. For example, the level of the fluctuations caused by the spatially varying baryon–dark matter relative velocity (Fig. 1) depends directly on how rapidly the scattering cross-section varies with velocity. In this regard, it is important to note that while in this study we have focused on the global 21-cm signal, the corresponding 21-cm fluctuations are also (over most of the parameter space shown in Fig. 3) much larger than would be expected in the absence of baryon–dark matter scattering. A more advanced test of $\sigma(v) \propto v^{-4}$ would be a comparison with a measurement of the signal during the dark ages (Fig. 2), which would involve different velocities and would provide constraints independent of astrophysics. However, even the proposed Dark Ages Radio Explorer (DARE) satellite[54] is planned to perform measurements only at frequencies greater than 40 MHz, whereas the predicted absorption signal from the dark ages is expected at 10–30 MHz.

Dark-matter millicharge and other dark-matter models. We have assumed a relatively simple model for the dark matter and its interactions with ordinary matter. However, this model is only an illustration; the EDGES data suggest a low-velocity baryon–dark matter interaction, but do not yet determine the specific form of this interaction. If the measured signal is confirmed and 21-cm cosmology becomes a dark-matter detector, then a wide range of candidate dark-matter models will need to be re-assessed. Examples include models that describe the interaction of dark matter with electrons[55,56] and interactions (with either baryons or electrons) that depend on whether standard-model particles are free or bound within atoms, models in which the dark matter consists of multiple components so that only a fraction of the particles interact with baryons, and models in which dark matter also self-interacts[57]. We add here a few notes on these possibilities. Dark-matter interactions with electrons are currently constrained[56] to lower masses than for interactions with baryons. Also, the formation of atoms obviously affects the motion of electrons much more than that of the baryons. Thus, although an interaction with baryons may not change much at recombination (as long as the interaction is not electromagnetic), one with electrons would be more likely to change, in a complex and model-dependent way. Now, any interaction that is strong only with free protons or electrons decreases in proportion to the residual electron or proton fraction after cosmic recombination. This means that the effect of dark-matter scattering on baryon cooling during cosmic dawn is suppressed relative to the effect on the CMB by a factor of about 5,000. At a temperature of 5 K (indicated by the EDGES data at $z = 17.2$, as noted in the main text), the r.m.s. velocity of baryons is about $0.3\,\mathrm{km\,s^{-1}}$, which is lower than the velocity at cosmic recombination by a factor of about 25. For a v^{-4} model, this makes the cross-section larger by a factor of approximately 500,000. Still, the cooling takes time and is unlikely to be effective unless it begins at velocities well above $1\,\mathrm{km\,s^{-1}}$. Thus, this type of model might be ruled out by CMB observations, which are very constraining, even for the model that we assumed in this work (see the next section).

It is interesting to consider the specific dark-matter millicharge model[8,18,19,58,59] that naturally yields a v^{-4} cross-section. In this case, the dark matter has a (small) standard electric charge. Considering only the interaction with protons and the aforementioned suppression factor of 5,000, the minimum dark-matter charge (relative to that of the electron) that would be required to produce the cooling indicated by the EDGES data is $\varepsilon \approx 10^{-6}$. The interaction with electrons can also be considerable if the dark-matter mass is well below 1 GeV, which might allow substantial cooling with values of ε as low as about 10^{-8}. However, the entire range of relevant values may already be ruled out: owing to its interaction with Galactic magnetic fields, millicharge dark matter should have been evacuated from the Milky Way disk (and also blocked from re-entering) if[18,19]

$$10^{-11}\frac{m_\chi}{1\ \mathrm{GeV}} < \varepsilon < 3\times10^{-4}\left(\frac{m_\chi}{1\ \mathrm{GeV}}\right)^{1/2}$$

However, multiple measurements indicate a non-zero local dark-matter density near the Sun[60], roughly consistent with the expected density from the Milky Way's dark-matter halo. This contradicts the idea that the dark matter was evacuated from the disk and thus rules out the existence of millicharge dark matter over a wide range of parameters (we note that this argument has not been made previously). Furthermore, CMB constraints[8,19,59] yield a conservative upper limit of $\varepsilon \approx 4\times10^{-6}[m_\chi/(1\,\mathrm{GeV})]^{1/2}$. These observational constraints combine to exclude $\varepsilon > 10^{-11}[m_\chi/(1\,\mathrm{GeV})]$ at all relevant values of m_χ, and this excludes the possibility of substantial cooling at cosmic dawn. The dark-matter millicharge model may be similarly (but independently) ruled out on the basis of the effect that the interaction of dark matter with magnetic fields would have on the dark-matter distribution in galaxy clusters[61]. These constraints all rely on the presence of the electric millicharge directly, while if other properties are assumed (for example, related to the dark-matter annihilation properties), other stringent constraints would also apply[58,59].

Comparing cosmic dawn constraints to other limits on baryon–dark matter interactions. The comparison between cosmic dawn observations for dark-matter detection and constraints from direct detection, accelerators and various astrophysical phenomena is model-dependent. Here we adopt the $\sigma(v) \propto v^{-4}$ model, in which case the parameter spaces of the various approaches overlap, and other searches may be able to detect or rule out a dark-matter particle that is consistent with the 21-cm observations at cosmic dawn. However, a more complex interaction—for example, based on a bound state or resonance that is important only at low velocities—could invalidate any such comparison and make cosmic dawn observations a unique probe. Additional model dependence enters some of the comparisons that involve assumptions about dark-matter annihilation or the spin dependence of the baryon–dark matter scattering.

Limits on the v^{-4} model have been derived previously[19], with the strongest limits based on CMB observations (plus a slight improvement from including clustering based on Lyman-α forest data)[8]. However, the 95% confidence limit equivalent to $\sigma_1 < 2\times10^{-19}[m_\chi/(1\,\mathrm{GeV})]\,\mathrm{cm}^2$ was derived only for $m_\chi \gg m_{\mathrm{H}}$. This calculation must be re-done for lower m_χ and fixed more generally to properly include the spatial variation of the baryon–dark matter relative velocity, which would introduce a CMB pattern that may be partially correlated with the standard one. Here we estimate a very rough correction for low dark-matter masses. In the limit of strong coupling (so that the gas and dark matter have the common temperature of equation (3)), including the contribution (neglected in the above limit) of the dark matter to the relative thermal velocity and assuming that the limit is proportional to v^{-4} gives a modified limit of $\sigma_1 < 2\times10^{-19}(m_\chi/\mathrm{GeV}) \times [1+(\mu_b/m_\chi)]^2\,\mathrm{cm}^2$. Nevertheless, if the coupling is not strong, then $T_\chi < T_{\mathrm{gas}}$ and the correction factor may be smaller. We conclude that CMB limits may complement the 21-cm signal by imposing useful upper limits on σ_1, but these limits must be carefully re-calculated. There is also a limit on the baryon–dark matter scattering from spectral distortions of the CMB[55], but these distortions occur at rather high redshifts (and thus high velocities) so are probably not important (and have not been considered) for a cross-section that peaks at low velocities, such as that of the $\sigma(v) \propto v^{-4}$ model.

Another possibility (related to the earlier discussion about the dark-matter velocity dispersion) is that between matter–radiation equality and recombination—a period when fluctuations normally grow in the dark matter but not the baryons—the baryon–dark matter coupling might suppress the growth of fluctuations in the dark matter; this effect could be consistent with CMB observations but also important for the power spectrum at low redshift.

A different limit on baryon–dark matter interactions comes from experiments that attempt to detect dark-matter scattering directly with target nuclei in the laboratory. Assuming typical Milky Way halo speeds, with velocities of about $200\,\mathrm{km\,s^{-1}}$, the minimum cross-section $\sigma_1 = 3.4\times10^{-21}\,\mathrm{cm}^2$ required for the cosmological 21-cm effect (Fig. 3) translates (in a v^{-4} model) to $\sigma(200\,\mathrm{km\,s^{-1}}) > 2\times10^{-30}\,\mathrm{cm}^2$. This is in the range of cross-sections that are hard to probe with underground detection experiments because at such cross-sections dark-matter particles are expected to lose most of their energy in the Earth's crust before reaching the detector[62]. The Cryogenic Rare Event Search with Superconducting Thermometers (CRESST) underground experiment is most relevant to the 21-cm parameter region, although it constrains only relatively high masses. At $m_\chi = 1$–$5\,\mathrm{GeV}$, it rules out (assuming spin-independent interactions throughout this discussion) values of $\sigma(200\,\mathrm{km\,s^{-1}})$ between $10^{-37}\,\mathrm{cm}^2$ and $(2$–$3) \times 10^{-31}\,\mathrm{cm}^2$ (the upper limit varies with m_χ), although the limit might change when re-calculated for a v^{-4} model (because the particles slow down as they scatter within the Earth). Therefore, experiments above the Earth's surface are more advantageous. A 1987 balloon experiment has ruled out $m_\chi > 2$–$3\,\mathrm{GeV}$ (the precise limit depends on uncertainties in the velocity distribution of the dark-matter halo)[62,63]. The rocket-based X-ray Quantum Calorimetry (XQC) experiment has excluded[64] $\sigma(200\,\mathrm{km\,s^{-1}}) > 1 \times 10^{-29}\,\mathrm{cm}^2$ for $m_\chi > 0.5\,\mathrm{GeV}$. The limits from all such experiments on lower dark-matter particle masses are quite weak, although this could change with new techniques[65,66].

A much stronger limit comes from the flip-side of the mentioned Earth interactions. The scattering of dark-matter particles within the Earth would heat it too strongly[67,68] unless $\sigma(200\,\mathrm{km\,s^{-1}}) < 10^{-32}(\mathrm{GeV}/m_\chi)\,\mathrm{cm}^2$, which is valid for m_χ values up to a few GeV; this implies that explaining the observed 21-cm signal at cosmic dawn requires $m_\chi < 5\,\mathrm{MeV}$. However, the Earth-heating constraint relies on some assumptions regarding dark-matter annihilation. Also, this and all the above direct detection limits on σ would be 3–4 orders of magnitude lower for spin-dependent interactions[62,68], whereas cosmic scattering with hydrogen would remain just as strong for such an interaction, given hydrogen's nuclear spin of $I = 1/2$ (note that ^{4}He would not contribute). In the case of a spin-dependent interaction, there might also be a direct effect of the dark-matter interaction on the spin temperature of hydrogen.

Because the collision energies in high-energy particle accelerators are typically orders of magnitude greater than 1 GeV, we assume that the relevant cross-section is σ_c. Accelerators may probe some of the parameter space that is relevant for 21-cm cosmology, but the limits depend strongly on the precise interaction type and nature of the dark-matter particle[66,69]. The proposed Search for Hidden Particles (SHiP) experiment at CERN has been motivated by the many possible physical mechanisms for producing very weakly interacting dark-matter particles in the MeV–GeV mass range[69]; cosmology may now provide additional impetus.

Astrophysical constraints on baryon–dark matter interactions are generally weaker than those we have considered here[67]. The most important limit, from cosmic rays, is often quoted as $\sigma(200\,\mathrm{km\,s^{-1}}) < 8 \times 10^{-27}[m_\chi/(1\,\mathrm{GeV})]\,\mathrm{cm}^2$, but this is valid only for large m_χ values (greater than about 100 GeV), and the limits on dark-matter particle masses in our range of interest are far weaker[70].

We again emphasize that we have assumed a v^{-4} dependence of the baryon–dark matter scattering cross-section, and any modification of this relation would have a major effect on the above comparisons of the cosmological signal with the various limits on baryon–dark matter interactions (as would other extensions of the parameter space, such as allowing only a fraction of the dark matter to scatter with baryons). Also, previous constraints have often been derived for a velocity-independent cross-section and must be carefully re-assessed for the case of a strong velocity dependence. The constraints on the dark-matter particle derived from the condition that it has the required relic cosmic density are model-dependent, as they depend on the annihilation cross-section (for thermal production) or the detailed production mechanism (for non-thermal production). If we assume an annihilation cross-section as expected for the weak interaction and a similar baryon–dark matter cross-section at relativistic velocities (so that $\sigma_c \approx 10^{-36}\,\mathrm{cm}^{-2}$), then the 21-cm signal suggests a dependence closer to v^{-3} than to v^{-4}.

Finally, we note that after this Letter was submitted, limits on baryon–dark matter scattering were derived[71] from low-redshift observations of the temperature of the intergalactic medium, based on the Lyman-α forest at $z \approx 5$. These limits are at relative velocities of about $10\,\mathrm{km\,s^{-1}}$, similar to the CMB limits discussed above. The derived upper limit of $\sigma_c = 3 \times 10^{-38}\,\mathrm{cm}^2$ for a v^{-4} model (for $m_\chi \ll 1\,\mathrm{GeV}$) is about five orders of magnitude greater than the minimum cross-section implied by the 21-cm signal at cosmic dawn, and it is stronger than the above CMB limit only for m_χ lower than about 1 MeV (although, as noted, the CMB limit must be carefully revised). Also, the low-redshift limit is uncertain because it depends on the history of photoheating of the intergalactic medium; the latter depends on the spatial and temporal distribution of the spectrum of ionizing sources, the distribution of Lyman-limit absorbers and the after-effects of inhomogeneous reionization, all of which are incompletely known. In the low-redshift probe, astrophysical heating is partly degenerate with baryon–dark matter interactions, and there is no unambiguous sign of dark matter similar to the excess absorption signal during cosmic dawn.

Code and data availability. We have opted not to make the code available because the calculations are based on the combination of published results on baryon–dark matter scattering[10] with our cosmic dawn numerical simulation code, the details and advantages of which have been described previously[15]. Confirmation of our basic results can be achieved by modifying other cosmic dawn codes, including publicly available ones[72]. The datasets generated or analysed in this study are available from the corresponding author on reasonable request.

31. Cohen, A., Fialkov, A., Barkana, R. & Lotem, M. Charting the parameter space of the global 21-cm signal. *Mon. Not. R. Astron. Soc.* **472,** 1915–1931 (2017).
32. Barkana, R. & Loeb, A. Detecting the earliest galaxies through two new sources of 21 centimeter fluctuations. *Astrophys. J.* **626,** 1–11 (2005).
33. Pritchard, J. R. & Furlanetto, S. 21-cm fluctuations from inhomogeneous X-ray heating before reionization. *Mon. Not. R. Astron. Soc.* **376,** 1680–1694 (2007).
34. Planck Collaboration. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594,** A13 (2016).
35. Ali-Haïmoud, Y. & Hirata, C. M. HyRec: a fast and highly accurate primordial hydrogen and helium recombination code. *Phys. Rev. D* **83,** 043513 (2011).
36. McQuinn, M. The evolution of the intergalactic medium. *Annu. Rev. Astron. Astrophys.* **54,** 313–362 (2016).
37. Miralda-Escudé, J., Haehnelt, M. & Rees, M. J. Reionization of the inhomogeneous Universe. *Astrophys. J.* **530,** 1–16 (2000).
38. Ross, H. E., Dixon, K. L., Iliev, I. T. & Mellema, G. Simulating the impact of X-ray heating during the cosmic dawn. *Mon. Not. R. Astron. Soc.* **468,** 3785–3797 (2017).
39. O'Leary, R. M. & McQuinn, M. The formation of the first cosmic structures and the physics of the $z \sim 20$ Universe. *Astrophys. J.* **760,** 4 (2012).
40. Semelin, B., Eames, E., Bolgar, F. & Caillat, M. 21SSD: a public data base of simulated 21-cm signals from the epoch of reionization. *Mon. Not. R. Astron. Soc.* **472,** 4508–4520 (2017).
41. Gnedin, N. Y. & Hui, L. Probing the Universe with the Lyα forest – I. Hydrodynamics of the low-density intergalactic medium. *Mon. Not. R. Astron. Soc.* **296,** 44–55 (1998).
42. Naoz, S. & Barkana, R. The formation and gas content of high-redshift galaxies and minihaloes. *Mon. Not. R. Astron. Soc.* **377,** 667–676 (2007).
43. Monsalve, R. A., Rogers, A. E. E., Bowman, J. D. & Mozdzen, T. J. Results from EDGES High-band. I. Constraints on phenomenological models for the global 21 cm signal. *Astrophys. J.* **847,** 64 (2017).
44. Ali, Z. S. *et al.* PAPER-64 constraints on reionization: the 21 cm power spectrum at $z = 8.4$. *Astrophys. J.* **809,** 61 (2015).
45. Patil, A. H. *et al.* Upper limits on the 21 cm epoch of reionization power spectrum from one night with LOFAR. *Astrophys. J.* **838,** 65 (2017).
46. Beardsley, A. P. *et al.* First season MWA EoR power spectrum results at redshift 7. *Astrophys. J.* **833,** 102 (2016).
47. Paciga, G. *et al.* A simulation-calibrated limit on the H I power spectrum from the GMRT Epoch of Reionization experiment. *Mon. Not. R. Astron. Soc.* **433,** 639–647 (2013).
48. Chuzhoy, L. & Shapiro, P. R. Ultraviolet pumping of hyperfine transitions in the light elements, with application to 21 cm hydrogen and 92 cm deuterium lines from the early universe. *Astrophys. J.* **651,** 1–7 (2006).
49. Chen, X. & Miralda-Escudé, J. The spin-kinetic temperature coupling and the heating rate due to Lyα scattering before reionization: predictions for 21 centimeter emission and absorption. *Astrophys. J.* **602,** 1–11 (2004).
50. Hirata, C. M. Wouthuysen–Field coupling strength and application to high-redshift 21-cm radiation. *Mon. Not. R. Astron. Soc.* **367,** 259–274 (2006).
51. Furlanetto, S. R. & Pritchard, J. R. The scattering of Lyman-series photons in the intergalactic medium. *Mon. Not. R. Astron. Soc.* **372,** 1093–1103 (2006).
52. Chuzhoy, L. & Shapiro, P. R. Heating and cooling of the early intergalactic medium by resonance photons. *Astrophys. J.* **655,** 843–846 (2007).
53. Hui, L., Ostriker, J. P., Tremaine, S. & Witten, E. Ultralight scalars as cosmological dark matter. *Phys. Rev. D* **95,** 043541 (2017).
54. Burns, J. O. *et al.* Probing the first stars and black holes in the early Universe with the Dark Ages Radio Explorer (DARE). *Adv. Space Res.* **49,** 433–450 (2012).
55. Ali-Haïmoud, Y., Chluba, J. & Kamionkowski, M. Constraints on dark matter interactions with standard model particles from cosmic microwave background spectral distortions. *Phys. Rev. Lett.* **115,** 071304 (2015).
56. Essig, R., Volansky, T. & Yu, T.-T. New constraints and prospects for sub-GeV dark matter scattering off electrons in xenon. *Phys. Rev. D* **96,** 043017 (2017).
57. Spergel, D. N. & Steinhardt, P. J. Observational evidence for self-interacting cold dark matter. *Phys. Rev. Lett.* **84,** 3760–3763 (2000).
58. Davidson, S., Hannestad, S. & Raffelt, G. Updated bounds on milli-charged particles. *J. High Energy Phys.* **5,** 3 (2000).
59. Dubovsky, S. L., Gorbunov, D. S. & Rubtsov, G. I. Narrowing the window for millicharged particles by CMB anisotropy. *J. Exp. Theor. Phys.* **79,** 1–5 (2004).
60. Read, J. I. The local dark matter density. *J. Phys. G* **41,** 063101 (2014).
61. Kadota, K., Sekiguchi, T. & Tashiro, H. A new constraint on millicharged dark matter from galaxy clusters. Preprint at https://arxiv.org/abs/1602.04009 (2016).
62. Zaharijas, G. & Farrar, G. R. Window in the dark matter exclusion limits. *Phys. Rev. D* **72,** 083502 (2005).
63. Rich, J., Rocchia, R. & Spiro, M. A search for strongly interacting dark matter. *Phys. Lett. B* **194,** 173–176 (1987).
64. Erickcek, A. L., Steinhardt, P. J., McCammon, D. & McGuire, P. C. Constraints on the interactions between dark matter and baryons from the x-ray quantum calorimetry experiment. *Phys. Rev. D* **76,** 042007 (2007).
65. Budnik, R., Chesnovsky, O., Slone, O. & Volansky, T. Direct detection of light dark matter and solar neutrinos via color center production in crystals. Preprint at https://arxiv.org/abs/1705.03016 (2017).

66. Battaglieri, M. *et al.* US cosmic visions: new ideas in dark matter 2017: community report. Preprint at https://arxiv.org/abs/1707.04591 (2017).
67. Starkman, G. D., Gould, A., Esmailzadeh, R. & Dimopoulos, S. Opening the window on strongly interacting dark matter. *Phys. Rev. D* **41,** 3594–3603 (1990).
68. Mack, G. D., Beacom, J. F. & Bertone, G. Towards closing the window on strongly interacting dark matter: far-reaching constraints from Earth's heat flow. *Phys. Rev. D* **76,** 043523 (2007).
69. Alekhin, S. *et al.* A facility to search for hidden particles at the CERN SPS: the SHiP physics case. *Rep. Prog. Phys.* **79,** 124201 (2016).
70. Cyburt, R. H., Fields, B. D., Pavlidou, V. & Wandelt, B. Constraining strong baryon-dark-matter interactions with primordial nucleosynthesis and cosmic rays. *Phys. Rev. D* **65,** 123503 (2002).
71. Muñoz, J. B. & Loeb, A. Constraints on dark-matter-baryon scattering from the temperature evolution of the intergalactic medium. *J. Cosmol. Astropart. Phys.* **2017,** 043 (2017).
72. Mesinger, A., Furlanetto, S. & Cen, R. 21CMFAST: a fast, seminumerical simulation of the high-redshift 21-cm signal. *Mon. Not. R. Astron. Soc.* **411,** 955–972 (2011).

doi:10.1038/nature25737

Fast quantum logic gates with trapped-ion qubits

V. M. Schäfer[1], C. J. Ballance[1], K. Thirumalai[1], L. J. Stephenson[1], T. G. Ballance[1], A. M. Steane[1] & D. M. Lucas[1]

Quantum bits (qubits) based on individual trapped atomic ions are a promising technology for building a quantum computer[1]. The elementary operations necessary to do so have been achieved with the required precision for some error-correction schemes[2–4]. However, the essential two-qubit logic gate that is used to generate quantum entanglement has hitherto always been performed in an adiabatic regime (in which the gate is slow compared with the characteristic motional frequencies of the ions in the trap[3–7]), resulting in logic speeds of the order of 10 kilohertz. There have been numerous proposals of methods for performing gates faster than this natural 'speed limit' of the trap[8–12]. Here we implement one such method[11], which uses amplitude-shaped laser pulses to drive the motion of the ions along trajectories designed so that the gate operation is insensitive to the optical phase of the pulses. This enables fast (megahertz-rate) quantum logic that is robust to fluctuations in the optical phase, which would otherwise be an important source of experimental error. We demonstrate entanglement generation for gate times as short as 480 nanoseconds—less than a single oscillation period of an ion in the trap and eight orders of magnitude shorter than the memory coherence time measured in similar calcium-43 hyperfine qubits. The power of the method is most evident at intermediate timescales, at which it yields a gate error more than ten times lower than can be attained using conventional techniques; for example, we achieve a 1.6-microsecond-duration gate with a fidelity of 99.8 per cent. Faster and higher-fidelity gates are possible at the cost of greater laser intensity. The method requires only a single amplitude-shaped pulse and one pair of beams derived from a continuous-wave laser. It offers the prospect of combining the unrivalled coherence properties[2,13], operation fidelities[2–4] and optical connectivity[14] of trapped-ion qubits with the submicrosecond logic speeds that are usually associated with solid-state devices[15,16].

Deterministic entanglement of multiple qubits—an essential prerequisite for general quantum information processing—was first achieved nearly twenty years ago using laser manipulation of qubits stored in the hyperfine ground states of trapped atomic ions[5]. Since then, progress in trapped-ion techniques, the development of more robust methods and an improved understanding of error sources have yielded a steady improvement in the precision of the fundamental two-qubit quantum logic gate, with the gate error ε_g decreasing by a factor of approximately two every two years, reaching $\varepsilon_g \approx 0.1\%$ in recent experiments[3,4]. In addition, all elementary single-qubit operations have been demonstrated with errors of less than 0.1% (refs 2–4). These error levels are already an order of magnitude below the threshold required for fault-tolerant quantum error-correction schemes[17]. By contrast, the two-qubit gate speed has remained fairly constant since the first demonstrations; the gates with the lowest reported errors had durations of 30 μs (ref. 4) and 100 μs (ref. 3). For qubits based on solid-state platforms, the interactions are much stronger, enabling much faster two-qubit operations (typically about 50 ns for superconducting circuits[15]; 480 ns for a recently demonstrated gate in silicon-based qubits[16]), but also leading to much shorter qubit coherence times (T_2^*; typically about 100 μs compared with about 1 min for atomic systems). Substantial progress has also been made in demonstrating simple algorithms and quantum simulations, which involve around 10 qubits, and in developing technologies that are amenable to scaling to larger numbers of qubits[18,19].

In previous trapped-ion experiments, the speed of the two-qubit gate operation has been limited by the use of methods that operate in an adiabatic regime with respect to the secular motional frequencies of the ions. Because these motional frequencies are typically about 1 MHz, gate durations are generally much longer than 1 μs, and attempts to increase the gate speed have resulted in larger gate errors; for example, $\varepsilon_g = 3\%$ was observed for the shortest reported gate time of $t_g = 5.3\,\mu s$ (ref. 3). (The minimum gate time given in ref. 3 (3.8 μs) is the full-width at half-maximum of the laser pulse, but for fair comparison with the gates reported here we quote the total gate time from the start of the rising edge to the end of the falling edge.) With recent progress in techniques for faster ground-state laser cooling[20], ion shuttling[21,22] and qubit readout[23], current two-qubit gate speeds threaten to be the limiting factor in the clock speed of a trapped-ion processor based on a quantum charge-coupled device (CCD) architecture[1,24], especially given that error-correction circuits typically contain more gates than

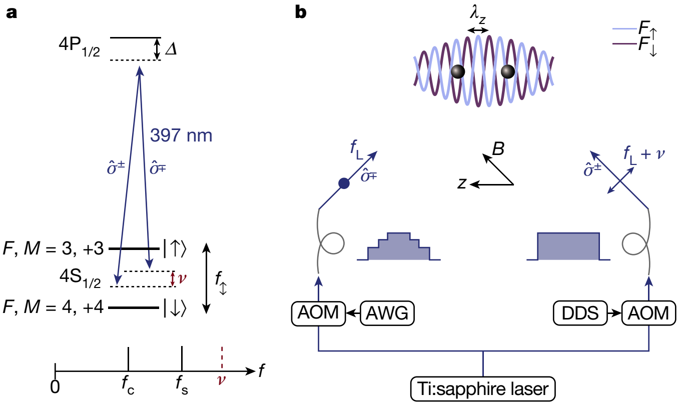


Figure 1 | Qubit states and Raman beam geometry. a, Qubits are stored in the $^{43}Ca^+$ hyperfine states $|\downarrow\rangle = 4S^{4,+4}_{1/2}$ and $|\uparrow\rangle = 4S^{3,+3}_{1/2}$, with separation $f_{\updownarrow} = 2.87$ GHz. The axial motional frequencies of the ion are $f_c = 1.92$ MHz and $f_s = 3.33$ MHz. The Raman beams (wavelength, 397 nm) have a mean detuning of $\Delta = -200$ GHz and a frequency difference of $\nu = 3.43 f_c$ for the fastest gate; for the highest-fidelity gates, $\Delta = -800$ GHz and $f_c < \nu < f_s$. **b**, The Raman beams of frequency f_L are derived from a single Ti:sapphire laser. One Raman beam propagates parallel to the quantization axis, set by a magnetic field $B \approx 14.6$ mT. The beams are perpendicular, such that their difference $\boldsymbol{k}$ vector is parallel to the z axis of the trap, and have waists of approximately 35 μm at the ions, powers of up to 200 mW and orthogonal linear polarizations $\hat{\sigma}^{\pm}$ and $\hat{\sigma}^{\mp}$. Their interference creates a polarization 'travelling standing wave' (period, $\lambda_z \approx 397\text{ nm}/\sqrt{2}$) that induces a spin-dependent force ($F_\downarrow$ or $F_\uparrow$) on the ions. High-bandwidth acousto-optic modulators (AOMs) shape the laser pulses on timescales of about 10 ns; we use a constant-amplitude pulse for one beam (right) and an amplitude-shaped pulse for the other beam (left). The AOMs are driven by an arbitrary-waveform generator (AWG) and a direct digital synthesis (DDS) source.

[1]Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK.

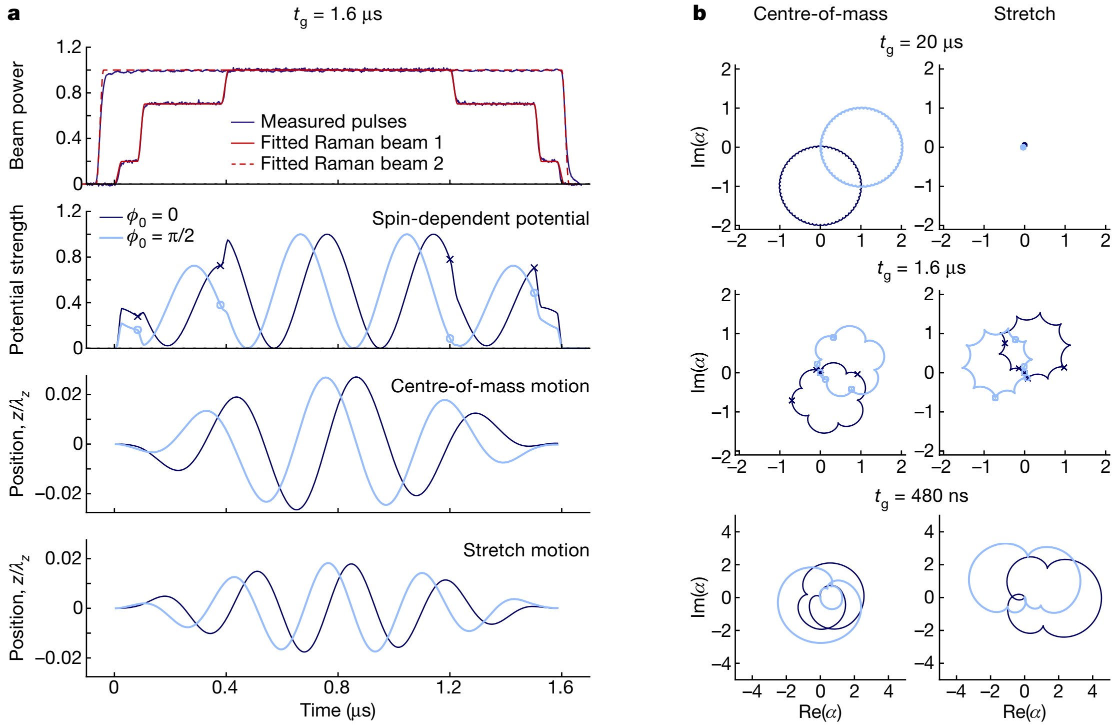


Figure 2 | Optical beat notes and motional trajectories of the ions for two initial optical phases ($\phi_0 = 0$ and $\phi_0 = \pi/2$). a, For the 1.6-µs-duration gate, the plots show: the Raman laser pulses (top); their calculated optical beat note, which gives rise to the spin- and position-dependent potential (and hence force) that the ions experience; the centre-of-mass displacement of the ions; and their stretch-mode displacement (bottom). The beat frequency is $\nu = 2.63$ MHz $\approx 1.37 f_c$. The force and motions clearly depend on ϕ_0; however, the shape of the pulse is designed so that, for all ϕ_0, both trajectories return to zero displacement at $t = t_g$. **b**, Phase-space trajectories for gates in three regimes (where α is the displacement in the rotating frame). For a conventional adiabatic gate ($t_g = 20$ µs; top), $\nu \approx 1.03 f_c$ and the stretch mode is barely excited; ϕ_0 affects the orientation of the (nearly circular) trajectory, but not its shape or area. For $t_g = 1.6$ µs (middle), both modes are driven and ϕ_0 affects the shape of the trajectories slightly; amplitude shaping is necessary to close the loops for both modes and to ensure that the net gate phase is independent of ϕ_0. Symbols correspond to steps in the pulse amplitude. For $t_g = 480$ ns $\approx 0.90/f_c$ (bottom), the trajectory depends strongly on ϕ_0; this simulation uses the Lamb–Dicke approximation, but the breakdown of this approximation leads to substantial gate errors because the loops no longer close.

state-preparation and readout operations. The quantum CCD architecture is a natural choice for implementing surface-code error-correction methods[17], although such methods can also be mapped onto one-dimensional ion chains. Errors due to ambient heating of the motion of the ions are proportional to t_g and will therefore be suppressed for fast gates; this is advantageous for microfabricated traps in which the ions are confined near to electrode surfaces and hence subject to greater electric field noise[25]. Spin-dephasing errors due to, for example, magnetic-field fluctuations (which typically have a $1/f$ noise spectrum), will likewise be reduced, allowing the use of qubit states that have first-order sensitivity to the magnetic field[26] (at least during gate operations, as is the case here).

The 'speed limit' set by the trap frequency f_c is not a fundamental barrier: the Coulomb interaction that is responsible for coupling the ions is almost instantaneous at the typical separation distance of trapped ions (3.5 µm here), and there have been various theoretical proposals for fast gates, with t_g less than about $1/f_c$ (see, for example, refs 8–12). None of these proposals has so far been realized, although a demonstration of the method proposed in ref. 8 has recently been reported[27], with $t_g = 18.5$ µs $\approx 23/f_c$. Here, after first exploring the limits of the conventional $\sigma_z \otimes \sigma_z$ gate mechanism[6] (where σ_z is the Pauli operator), we implement the scheme proposed in ref. 11, in which the single rectangular laser pulse used in the conventional adiabatic method is replaced by a pulse whose amplitude is shaped in time.

The operation of the gate relies on a qubit-state-dependent force, which originates from a spatially varying light shift. This shift is caused by a 'travelling standing wave', which is generated by the optical interference pattern of two non-copropagating laser beams with frequencies that differ by ν (Fig. 1). We focus on the case of two ions, with the force coupling to only the axial modes of motion. We discuss the behaviour in three regimes: (1) a single rectangular pulse in the adiabatic regime, (2) a single rectangular pulse in the non-adiabatic regime and (3) a fast, shaped pulse or pulses.

In the first regime, by choosing $\nu = f_c + \delta$ with $\delta \ll f_c$, only the centre-of-mass normal mode at frequency f_c is excited (to a first approximation) and the rotating-wave approximation holds for the treatment of the motion. Starting from a state cooled to the Lamb–Dicke regime ($\eta^2 n \ll 1$, where η is the Lamb–Dicke parameter and n is the motional quantum number), the motion traces out an approximately circular path in the (rotating-frame) phase space of the harmonic oscillator, returning to its starting point after time $t_g = 1/\delta$ (Fig. 2). The geometric gate phase Φ is determined by the (signed) area enclosed by this path, which is proportional to Ω^2, where Ω is the Rabi frequency. We require $\Phi = \pi/2$ to generate the maximally entangled state $(|\downarrow\downarrow\rangle + i|\downarrow\uparrow\rangle + i|\uparrow\downarrow\rangle + |\uparrow\uparrow\rangle)/2$ from the separable state $(|\downarrow\downarrow\rangle + |\downarrow\uparrow\rangle + |\uparrow\downarrow\rangle + |\uparrow\uparrow\rangle)/2$ after time t_g. The gate phase Φ is independent of both the initial motional state (within the Lamb–Dicke regime) and the phase ϕ_0 of the optical beat note at the start time $t = 0$. The latter is crucial for achieving high gate fidelity in the laboratory, because ϕ_0 is sensitive to nanometre-scale length differences between the two laser-beam paths. Such gates have been implemented previously[3,6].

In the second regime, the gate time is reduced by increasing δ, but for δ comparable to f_c there are three complicating factors. First, both the centre-of-mass mode and the stretch mode (at frequency $f_s = \sqrt{3} f_c$) of a two-ion crystal will be excited and the associated trajectories in

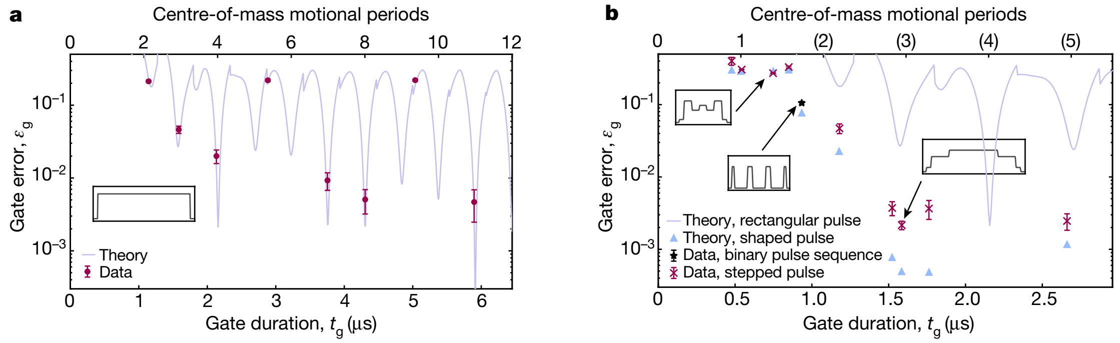


Figure 3 | Theoretical and experimental two-qubit gate errors. **a**, Conventional single, rectangular pulse (see inset). The curve shows the coherent error that is achievable (that is, excluding photon scattering and technical errors). At each time, Ω and ν are adjusted to minimize ε_g; discontinuities occur where the optimum value of ν switches between satisfying $f_c < \nu < f_s$ and $\nu > f_s$. Although a substantial reduction in gate error can be achieved by shaping the pulse edges[3] when $t_g \gg 1/f_c$, negligible improvement is possible when the gate duration becomes comparable to the shaping time constant, for $t_g \lesssim 4.5/f_c$. Data points show experimentally measured gate errors. **b**, Amplitude-shaped pulses. The values on the top axis given in parentheses are approximate. The curve from **a** is repeated for comparison (solid line). Simulated errors (triangles) are dominated by effects due to the breakdown of the Lamb–Dicke approximation for $t_g < 1.5$ μs. The other points (stars, binary pulses; crosses, stepped pulses) are gate errors measured after optimizing the pulse-shape parameters using real-time feedback from the experiment. Insets illustrate example pulse shapes. Error bars in **a** and **b** show 1σ statistical errors.

phase space will not in general close at the same time. Second, the trajectories depend on ϕ_0. Third, there is a time-dependent light shift that is independent of the motion but which also depends on ϕ_0 and can result in a large single-qubit phase ϕ_{LS}. Consequently, the expected gate error has a complicated dependence on gate time, and rises steeply as the gate time approaches the period of the motion. This behaviour is demonstrated in Fig. 3a, together with a selection of results from our experiments. We measure a gate error of $\varepsilon_g = 2.0(5)\%$ for $t_g = 2.13$ μs, and the theory shows that no solutions exist with lower errors at shorter times.

In the third regime, replacing the single rectangular pulse of the conventional method by a shaped pulse results in more degrees of freedom (the parameters that describe the shape of the pulse), which can be used to find especially well-performing or 'magic' pulses. In particular, we want to achieve all of the following: that the phase-space trajectories for both modes close simultaneously at $t = t_g$; that the sum of the enclosed (signed) areas is independent of ϕ_0, even though the trajectories themselves may depend on ϕ_0; that the light-shift-induced phase ϕ_{LS} is independent of ϕ_0 and preferably small; that the pulse area is small to minimize photon scattering[28]; and that the gate error is not too sensitive to errors in the parameter settings. A shaped pulse or pulse sequence is deemed a solution when it has all of these properties, and the gate error predicted for a perfectly realized sequence is below an upper bound ε_t set by practical considerations. In other words, we set ε_t well below the error that we are prepared to accept in the laboratory and search for solutions numerically.

Several classes of solution are given in ref. 11 for simple pulse shapes. We implement two types of time-symmetric sequence: a binary pulse sequence (in which a constant-amplitude force is switched on and off), and a five- or seven-segment stepped pulse. Example phase-space trajectories are shown in Fig. 2; we obtained gates up to an order of magnitude faster than those demonstrated previously. However, to understand the experimentally observed gate error and the optimal pulse shapes, we had to develop the theory further.

The solutions given in ref. 11 assume that the motion remains within the Lamb–Dicke regime. For t_g comparable to $1/f_c$ this is a poor approximation because large excursions in phase space are required to enclose sufficient area. The ions then become sensitive to the spatial variation in the force, leading to modification of the trajectories and squeezing of the motional wave packets[29]. We extend the theory by using numerical modelling to include the effects of this motional excursion beyond the Lamb–Dicke regime, and find solutions that give the minimum gate error for times in the range 200 ns $< t_g < 5.0$ μs (see Methods). The most efficient solutions, which involve optimal use of the available laser power, are found for $f_c < \nu < f_s$, when both modes are excited such that the geometric phases from each mode add constructively ($\Phi = \Phi_c + \Phi_s$). Conversely, when $\nu > f_s$, the phases subtract and more laser power is required to achieve $\Phi = \pi/2$, which in turn leads to higher photon scattering error[28]. The numerically calculated errors for some of these efficient solutions are shown in Fig. 3b, together with experimentally achieved gate errors for gate times between 480 ns and 2.7 μs (see Methods for experimental details). The fastest gate time is slightly below the centre-of-mass motional period ($1/f_c = 540$ ns), but the error is large (40%). The binary pulse sequence achieves 11% error at $t_g = 0.93$ μs. The minimum error of 0.22(3)% at $t_g = 1.6$ μs is obtained using a stepped pulse and is close to the lowest two-qubit gate errors reported previously[3,4], while the gate is 20–60 times faster. This error is an order of magnitude lower than that achievable using the conventional single-rectangular-pulse method at the same t_g. For the 1.6-μs gate, we estimate the total error due to known sources to be approximately 0.18% (Table 1).

In our set-up, the gate speed and fidelity are limited by the breakdown of the Lamb–Dicke approximation for $t_g \lesssim 1/f_c$. Faster and higher-fidelity gates are possible by reducing the Lamb–Dicke parameters (here $\eta_c = 0.126$ and $\eta_s = 0.096$); for example, decreasing the 90° angle between the two laser beams (Fig. 1b) to give $\eta_c = 0.08$ would reduce the error contribution due to the breakdown of the Lamb–Dicke approximation to 7×10^{-5}. Maintaining the same gate speed would then require higher laser intensities at the ions; although we use a moderately high laser power (about 150 mW per beam for the fastest gate), the intensity is modest (about 0.1 mW μm^{-2}) and the spot size ($w_0 \approx 35$ μm) could be reduced substantially. Alternatively, if the optical phase ϕ_0 could be controlled sufficiently, solutions could be found for fixed ϕ_0 that allow faster gates and higher fidelities[12].

Table 1 | Error budget for the highest-fidelity and fastest gates achieved

Error source	$t_g = 1.6$ μs	$t_g = 480$ ns
Lamb–Dicke-approximation breakdown	5×10^{-4}	3×10^{-1}
Optical phase chirp	$\approx 4 \times 10^{-4}$	$\approx 6 \times 10^{-3}$
Pulse timing and amplitudes	$\approx 2 \times 10^{-4}$	$\approx 1 \times 10^{-3}$
Radial mode excitation	$\lesssim 4 \times 10^{-5}$	$\lesssim 4 \times 10^{-3}$
Photon scattering	6×10^{-4}	7×10^{-3}
Centre-of-mass heating rate	8×10^{-5}	3×10^{-5}
Total error	1.8×10^{-3}	3.3×10^{-1}

The total error is the linear sum of the individual errors, on the assumption that they are constant and add incoherently.

We have demonstrated a method for realizing fast (1.6-μs duration) and robust two-qubit gates for trapped-ion qubits that combines state-of-the-art gate fidelity (99.8%) with more than an order of magnitude increase in gate speed compared to other methods. At the fastest speed demonstrated (480 ns), the fidelity achieved (60%) may not be useful for information processing, but might have other applications (such as quantum logic spectroscopy of short-lived exotic species[30,31], although this would require the use of fast laser-cooling techniques[32]). The method is technically simple, requiring only a single amplitude-shaped pulse from a continuous-wave laser, and the laser intensities required are within reach of miniature solid-state violet diodes[33]. These considerations are important if the techniques are ultimately to be scaled to the very large numbers of qubits necessary for an error-corrected quantum computer.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 October 2017; accepted 9 January 2018.

1. Wineland, D. J. *et al.* Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl Inst. Stand. Technol.* **103,** 259–328 (1998).
2. Harty, T. P. *et al.* High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit. *Phys. Rev. Lett.* **113,** 220501 (2014).
3. Ballance, C. J. *et al.* High-fidelity quantum logic gates using trapped-ion hyperfine qubits. *Phys. Rev. Lett.* **117,** 060504 (2016).
4. Gaebler, J. P. *et al.* High-fidelity universal gate set for ^{9}Be$^+$ ion qubits. *Phys. Rev. Lett.* **117,** 060505 (2016).
5. Turchette, Q. A. *et al.* Deterministic entanglement of two trapped ions. *Phys. Rev. Lett.* **81,** 3631–3634 (1998).
6. Leibfried, D. *et al.* Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. *Nature* **422,** 412–415 (2003).
7. Benhelm, J. *et al.* Towards fault-tolerant quantum computing with trapped ions. *Nat. Phys.* **4,** 463–466 (2008).
8. García-Ripoll, J. J., Zoller, P. & Cirac, J. I. Speed optimized two-qubit gates with laser coherent control techniques for ion trap quantum computing. *Phys. Rev. Lett.* **91,** 157901 (2003).
9. Duan, L.-M. Scaling ion trap quantum computation through fast quantum gates. *Phys. Rev. Lett.* **93,** 100502 (2004).
10. García-Ripoll, J. J., Zoller, P. & Cirac, J. I. Coherent control of trapped ions using off-resonant lasers. *Phys. Rev. A* **71,** 062309 (2005).
11. Steane, A. M. *et al.* Pulsed force sequences for fast phase-insensitive quantum gates in trapped ions. *New J. Phys.* **16,** 053049 (2014).
12. Palmero, M. *et al.* Fast phase gates with trapped ions. *Phys. Rev. A* **95,** 022328 (2017).
13. Wang, Y. *et al.* Single-qubit quantum memory exceeding ten-minute coherence time. *Nat. Photon.* **11,** 646–650 (2017).
14. Moehring, D. L. *et al.* Entanglement of single-atom quantum bits at a distance. *Nature* **449,** 68–71 (2007).
15. Barends, R. *et al.* Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature* **508,** 500–503 (2014).
16. Veldhorst, M. *et al.* A two-qubit logic gate in silicon. *Nature* **526,** 410–414 (2015).
17. Fowler, A. G. *et al.* Surface codes: towards practical large-scale quantum computation. *Phys. Rev. A* **86,** 032324 (2012).
18. Monroe, C. & Kim, J. Scaling the ion trap quantum processor. *Science* **339,** 1164–1169 (2013).
19. Devoret, M. H. & Schoelkopf, R. J. Superconducting circuits for quantum information: an outlook. *Science* **339,** 1169–1174 (2013).
20. Lin, Y. *et al.* Sympathetic electromagnetically-induced-transparency laser cooling of motional modes in an ion chain. *Phys. Rev. Lett.* **110,** 153002 (2013).
21. Bowler, R. *et al.* Coherent diabatic ion transport and separation in a multizone trap array. *Phys. Rev. Lett.* **109,** 080502 (2012).
22. Ruster, T. *et al.* Experimental realization of fast ion separation in segmented Paul traps. *Phys. Rev. A* **90,** 033410 (2014).
23. Noek, R. *et al.* High speed, high fidelity detection of an atomic hyperfine qubit. *Opt. Lett.* **38,** 4735–4738 (2013).
24. Kielpinski, D., Monroe, C. & Wineland, D. J. Architecture for a large-scale ion-trap quantum computer. *Nature* **417,** 709–711 (2002).
25. Turchette, Q. A. *et al.* Heating of trapped ions from the quantum ground state. *Phys. Rev. A* **61,** 063418 (2000).
26. Ruster, T. *et al.* A long-lived Zeeman trapped-ion qubit. *Appl. Phys. B* **122,** 254 (2016).
27. Wong-Campos, J. D., Moses, S. A., Johnson, K. G. & Monroe, C. Demonstration of two-atom entanglement with ultrafast optical pulses. *Phys. Rev. Lett.* **119,** 230501 (2017).
28. Ozeri, R. *et al.* Errors in trapped-ion quantum gates due to spontaneous photon scattering. *Phys. Rev. A* **75,** 042329 (2007).
29. McDonnell, M. J. *et al.* Long-lived mesoscopic entanglement outside the Lamb-Dicke regime. *Phys. Rev. Lett.* **98,** 063603 (2007).
30. Schmidt, P. O. *et al.* Spectroscopy using quantum logic. *Science* **309,** 749–752 (2005).
31. Meyer, V. *et al.* Measurement of the 1*s*–2*s* energy interval in muonium. *Phys. Rev. Lett.* **84,** 1136–1139 (2000).
32. Machnes, S. *et al.* Superfast laser cooling. *Phys. Rev. Lett.* **104,** 183001 (2010).
33. Schäfer, V. M. *et al.* Optical injection and spectral filtering of high-power ultraviolet laser diodes. *Opt. Lett.* **40,** 4265–4268 (2015).

Acknowledgements This work was supported by the UK EPSRC 'Networked Quantum Information Technologies' Hub, and the UK Defence, Science and Technology Laboratory. V.M.S. acknowledges funding from Balliol College, Oxford. C.J.B. acknowledges funding from Magdalen College, Oxford. We thank S. R. Woodrow for work on the trap design, T. P. Harty for contributions to the apparatus and W. Zhang for the loan of the AWG. We acknowledge the use of the University of Oxford Advanced Research Computing facility (https://doi.org/10.5281/zenodo.22558). The experiments benefitted from the use of the ARTIQ control system (https://doi.org/10.5281/zenodo.591804).

Author Contributions C.J.B. performed the numerical modelling. V.M.S. and C.J.B. designed and performed the experiments and analysed the data. K.T. built the ion trap and characterized the fast AOMs. L.J.S. and T.G.B. built optical and control systems. V.M.S., C.J.B., A.M.S. and D.M.L. wrote the manuscript, which all authors discussed.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.M.L. (d.lucas@physics.ox.ac.uk).

METHODS

Numerical modelling. Most trapped-ion experiments can be described in the Lamb–Dicke regime, in which the optical field is assumed to be uniform over the extent of the wavefunction of each ion. However, for the large phase-space displacements necessary to perform fast gates, this assumption breaks down—the curvature of the field can no longer be neglected. This means that the force experienced by an ion depends on its displacement in phase space, and this leads to squeezing of the wavefunction and modification of the motional trajectory.

To model the coherent error of a given gate sequence, we therefore numerically integrate the full Hamiltonian (without making the Lamb–Dicke approximation) using the split-operator method, explicitly averaging over different initial optical phases. Because this is a computationally intensive process, the gate sequences used in the experiments were preselected by an efficient solver that works in the Lamb–Dicke regime. Following ref. 11, we optimize candidate solutions starting from a random seed, and select a set of candidate solutions that have an error of less than 10^{-4} in the Lamb–Dicke approximation.

These candidate solutions were then evaluated using the full solver, and the most promising were optimized further. For the experiments, we chose solutions from this set by looking for a combination of low coherent error and low integrated pulse area (this both selects for a low photon scattering error[28] and avoids fragile sequences that use large motional excitations, which are more sensitive to parameter variations).

We evaluated several different pulse shapes. The seven-segment symmetric pulse shape offered a sufficient number of parameters to find a dense set of good solutions, while being easy to implement and to verify. The exact shape of the rising and falling edges is unimportant: the rise time can be varied from zero to the segment length without a change in gate fidelity, providing that an overall scaling factor is applied to the Rabi frequency of the gate to compensate for the changing spectral content.

Raman beams. The light source for the Raman beams is a frequency-doubled Ti:sapphire laser (M-Squared SolsTiS ECD-X) with 1.8-W output power at 397 nm. The Raman detuning was $\Delta = -1$ THz for single rectangular pulses; for shaped-pulse gates with $t_g \leq 1\,\mu$s, $\Delta = -200$ GHz; and for $t_g > 1\,\mu$s, $\Delta = -800$ GHz. The detuning was changed to reduce photon scattering errors for gates that require lower Rabi frequencies. For the fastest gate, peak powers of 192 mW and 96 mW were used for the two Raman beams, which had waists at the ions of 33 μm and 38 μm, respectively ($1/e^2$ intensity radius). The ratios of the Raman-beam intensities were chosen so that the scattering error was approximately minimized. The beams were modulated by a pair of acousto-optic modulators (AOMs; Brimrose CQM-200-40-.400/OW), with a 24-ns rise time (10%–90%) to create the shaped pulses that drive the gate. The amplitude-shaped radio-frequency (200-MHz) signal for the stepped pulse was defined using an arbitrary-waveform generator (AWG; Agilent N8241A, 1.25-GHz clock rate, 15 bits vertical resolution) and fed to the first AOM. The second AOM was driven by a direct digital synthesis (DDS; Enterpoint Milldown card, 200 MHz) source (Fig. 1b).

Phase chirps of the modulated beam were measured in an optical homodyne experiment and found to be substantial during switching of the radio-frequency amplitude. Driving the AOM at its centre frequency (200 MHz) minimized the phase chirps[34] so that their contribution to the gate error was small (Table 1).

Pulse calibration. Performing fast gates with high fidelities requires precise control of the pulse parameters. Owing to the nonlinear response of the AOM for different radio-frequency drive amplitudes, the pulse shape was measured on a photodiode and the relative drive amplitudes of each pulse segment were adjusted to match the measured amplitudes to their theoretically predicted optimum levels. The relative amplitudes of the stepped pulses were set with ±0.2% accuracy. Although the exact shape of the rising and falling edges is not critical, for a given shape the duration of the pulse segments needs to be set with subnanosecond precision. The waveform programmed into the 1.25 Gsps (1 Gsps = 10^9 samples per second) AWG had a 5.0-ns rise time, to spread the pulse edges over several time points and to improve the effective timing resolution. The timing precision of the optical pulses was measured to be 0.2 ns (standard deviation of fitted pulse lengths). Setting the pulse-shape parameters to their theoretically predicted values yielded optimal fidelity for all gate sequences for which the Lamb–Dicke approximation held well. There are three remaining parameters that characterize the gate sequence: the peak beam power, the Raman beat-note frequency ν and the phase offset $\phi_{\pi/2}$ of the last π/2 pulse of the Ramsey interferometer ($\phi_{\pi/2}$ compensates for the single-qubit phase that is acquired during the gate). The beat-note frequency and beam power were set to their theoretically predicted values and then optimized empirically; in all cases the optimized values agreed well with the theoretical predictions. The peak pulse powers of each beam were stabilized at the beginning of each experimental sequence. The phase offset $\phi_{\pi/2}$ was calibrated empirically. Initially gate parameters were optimized using a Nelder–Mead algorithm, using real-time feedback from the experiment. After the minimization of the optical phase chirps, this optimization method was no longer necessary and linear optimization of single parameters was found to be sufficient. A list of parameters for the fastest and highest-fidelity gates is given in Extended Data Table 1.

Experimental procedure. All gates were performed in a blade-type linear Paul trap[35,36], with axial centre-of-mass frequency $f_c = 1.92$ MHz for $t_g > 1\,\mu$s, and $f_c = 1.86$ MHz for $t_g \leq 1\,\mu$s and for all single-rectangular-pulse gates. The axial frequency was changed after re-aligning the Raman beams to suppress coupling to radial modes. In both cases the axial frequency was chosen so that the ion spacing was $12.5\lambda_z$, where $\lambda_z = 283$ nm is the periodicity of the travelling standing wave that provides the gate force. The gate was performed on the qubit states $|{\downarrow}\rangle = 4\mathrm{S}_{1/2}|F=4, M=+4\rangle$ and $|{\uparrow}\rangle = 4\mathrm{S}_{1/2}|F=3, M=+3\rangle$ in $^{43}\mathrm{Ca}^+$ at $B = 14.6$ mT. (This value of the B-field gives access to the 'atomic clock' qubit states $|{\downarrow'}\rangle = 4\mathrm{S}_{1/2}|F=4, M=0\rangle$ and $|{\uparrow'}\rangle = 4\mathrm{S}_{1/2}|F=3, M=+1\rangle$ with long coherence time T_2^*, measured previously[2] to be roughly 1 min, which are ideal for use as a memory qubit.) The ions were laser-cooled using dark-resonance Doppler cooling[37] to $\bar{n} \approx 1.8$ and further cooled using sideband cooling to $\bar{n} \lesssim 0.05$. After state preparation in $|{\downarrow\downarrow}\rangle$, we created an entangled state by placing the geometric phase gate in one arm of a Ramsey interferometer split by a spin-echo π pulse[6]. The gate errors were determined by using partial tomography[38] to measure the fidelity of the created state with respect to the desired state $(|{\downarrow\downarrow}\rangle + |{\uparrow\uparrow}\rangle)/\sqrt{2}$.

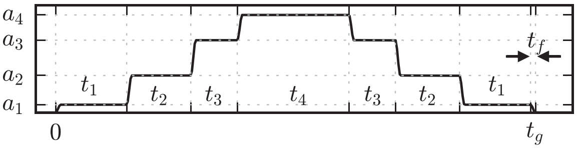
Error analysis. All gate errors and fidelities quoted are after correction for state-preparation and readout errors[3]. The total state-preparation and readout error with two ions was typically $\bar{\varepsilon}_{\mathrm{SPAM}} = 1.4(1) \times 10^{-3}$ per ion (averaged over both qubit states). The 0.22(3)% error reported for the $t_g = 1.6\,\mu$s gate is the average of five experimental runs measured over two days. Directly after calibrating the experimental parameters, the lowest error measured was 0.15(3)%; an hour after calibration, the measured error was 0.28(3)%. Quoted uncertainties are statistical only. We also measured the accumulated error for concatenated sequences of up to seven gates and found no evidence of coherent errors.

Errors due to radial-mode excitation are largest for gate times of $t_g \approx 800$ ns, because in this case the Raman beat-note frequency ν is close to resonance with the radial-mode frequencies (about 4.2 MHz). With the final alignment of the Raman laser beam we can limit errors due to radial-mode excitation to $\varepsilon_g < 5 \times 10^{-2}$ at $t_g = 800$ ns. An advantage of fast gates is that they are insensitive to errors associated with motional decoherence or heating: despite the relatively large heating rate of this trap ($\dot{\bar{n}} \approx 100\ \mathrm{s}^{-1}$ for the axial centre-of-mass mode), the contribution to the gate error is negligible. A summary of the main errors present in our experiments, for the lowest-error gate and for the fastest gate, is given in Table 1.

Data availability. The data shown in Figs 1–3 and that support the other findings of this study are available from the corresponding author on reasonable request.

34. Degenhardt, C. *et al.* Influence of chirped excitation pulses in an optical clock with ultracold calcium atoms. *IEEE Trans. Instrum. Meas.* **54,** 771–775 (2005).
35. Gulde, S. T. *Experimental Realization of Quantum Gates and the Deutsch–Josza Algorithm with Trapped ^{40}Ca$^+$-Ions.* PhD thesis, Univ. Innsbruck (2003).
36. Woodrow, S. R. *Linear Paul Trap Design for High-fidelity, Scalable Quantum Information Processing.* MSc thesis, Univ. Oxford (2015).
37. Allcock, D. T. C. *et al.* Dark-resonance Doppler cooling and high fluorescence in trapped Ca-43 ions at intermediate magnetic field. *New J. Phys.* **18,** 023043 (2016).
38. Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404,** 256–259 (2000).

Extended Data Table 1 | Gate parameters used for the fastest gate (seven segments) and for the highest-fidelity gate (five segments)



	gate duration	
parameter	$t_g = 483\,\text{ns}$	$t_g = 1.59\,\mu\text{s}$
Raman detuning Δ	$-200\,\text{GHz}$	$-800\,\text{GHz}$
Raman beat note frequency ν	6.3802 MHz	2.6301 MHz
axial centre-of-mass frequency f_c	1.8615 MHz	1.9243 MHz
peak power (pulse-shaped beam)	192 mW	58 mW
power (non-shaped beam)	96 mW	48 mW
single-qubit phase $\phi_{\pi/2}$	91.4°	21.4°
pulse time t_1	71.4 ns	82.1 ns
pulse time t_2	64.5 ns	299.9 ns
pulse time t_3	46.7 ns	—
pulse time t_4	112.3 ns	819.5 ns
pulse fall-time t_f	5.0 ns	5.0 ns
pulse amplitude a_1	0.284	0.445
pulse amplitude a_2	0.617	0.838
pulse amplitude a_3	0.862	—
pulse amplitude a_4	1	1

The pulse envelope (top) illustrates the definition of the pulse amplitude and timing parameters ($t_g = 2t_1 + 2t_2 + 2t_3 + t_4 + t_f$). The timing parameters refer to the timing of the waveform programmed into the AWG, for which a $t_f = 5.0$ ns rise/fall time (0%–100%) was used; the measured rise/fall time (10%–90%) of the laser pulses was 24 ns, owing to the bandwidth of the particular AOMs used (see Fig. 2a). The waists ($1/e^2$ intensity radii) of the Raman beams were 33 μm and 38 μm for the pulse-shaped and non-shaped beams, respectively.

LETTER

doi:10.1038/nature25484

Probing the interatomic potential of solids with strong-field nonlinear phononics

A. von Hoegen[1], R. Mankowsky[1], M. Fechner[1], M. Först[1] & A. Cavalleri[1,2]

Nonlinear optical techniques at visible frequencies have long been applied to condensed matter spectroscopy[1]. However, because many important excitations of solids are found at low energies, much can be gained from the extension of nonlinear optics to mid-infrared and terahertz frequencies[2,3]. For example, the nonlinear excitation of lattice vibrations has enabled the dynamic control of material functions[4–8]. So far it has only been possible to exploit second-order phonon nonlinearities[9] at terahertz field strengths near one million volts per centimetre. Here we achieve an order-of-magnitude increase in field strength and explore higher-order phonon nonlinearities. We excite up to five harmonics of the A_1 (transverse optical) phonon mode in the ferroelectric material lithium niobate. By using ultrashort mid-infrared laser pulses to drive the atoms far from their equilibrium positions, and measuring the large-amplitude atomic trajectories, we can sample the interatomic potential of lithium niobate, providing a benchmark for *ab initio* calculations for the material. Tomography of the energy surface by high-order nonlinear phononics could benefit many aspects of materials research, including the study of classical and quantum phase transitions.

In the experiments reported here, the highest-frequency A_1 mode of $LiNbO_3$ was excited with mid-infrared femtosecond pulses tuned to 17.5 THz, immediately to the red of the transverse-optical phonon frequency ($\nu_{TO} = 19$ THz)[10,11]. In the linear response regime, the real-space distortions of this mode involve rotations of the oxygen octahedra, accompanied by *c*-axis motions against the niobium and lithium sublattices (see Fig. 1a). Owing to the broken inversion symmetry of the crystal, the A_1 mode is both Raman- and infrared-active[10,11], with electric dipole moment along the *c* axis. Here, we explore the response of this mode up to very high amplitudes.

To study the dynamics of the driven mode, we measured time-dependent polarization rotation and second-harmonic intensity using 30-fs-long probe pulses at a wavelength of 800 nm. The polarization rotation yielded changes in the dielectric permittivity of the crystal $\varepsilon_r(\tau)$, whereas the second harmonic sampled the changes in the optical second-order susceptibility $\chi^{(2)}(\tau)$ (refs 12, 13) and with it the polar component of the lattice motion. Crucially, the stable absolute carrier-envelope phase (CEP)[14] of the pump field (Fig. 1b) made it possible to follow the atomic trajectories directly. Spectral interferometry between the polarization rotation and second-harmonic signals and their respective local oscillators derived from the same probe pulses yielded both the phase and the amplitude of these dynamics. The time resolution of these experiments was dictated by the bandwidths of the local oscillators on the detector[15,16], 60 THz and 80 THz for the second harmonic and polarization rotation, respectively. Hence, the measurements were sensitive to the phase of the signal oscillations up to the fifth overtone of the excited transverse-optical phonon mode (see Methods and Extended Data Figs 1–3 for details).

For small-amplitude excitation (0.1 MV cm^{-1}), both polarization rotation and second-harmonic measurements yielded harmonic oscillations (see Fig. 1c, d, dashed lines), which were readily attributed to a combination of a 15 THz phonon-polariton and the 19 THz transverse-optical phonon of the A_1 mode[17]. As shown in Methods (Extended Data Fig. 4), the pump–probe spectrum of the small-field response is well understood by considering the

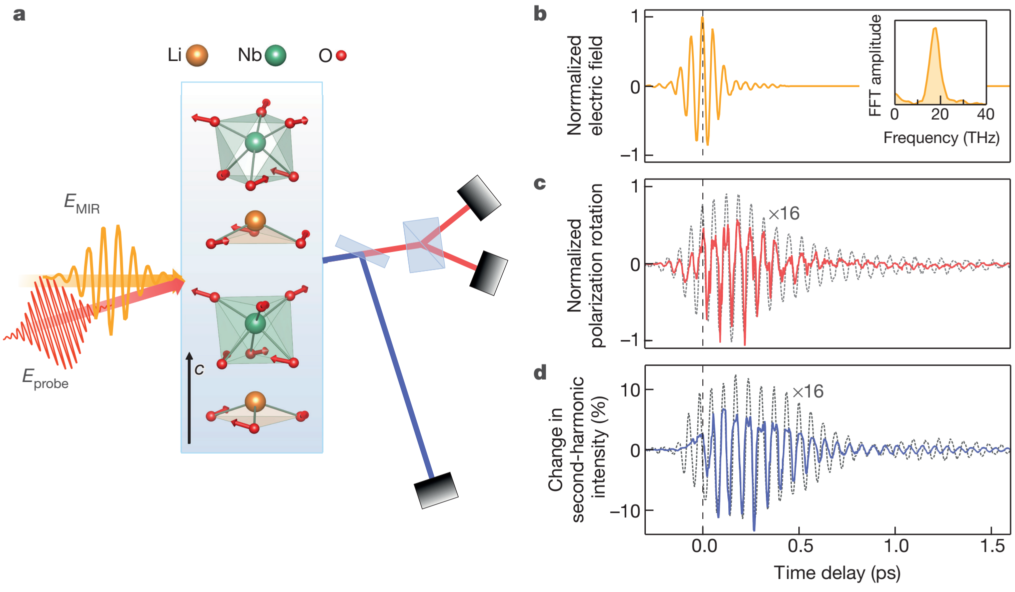


Figure 1 | Experimental set-up and time-resolved optical response. **a**, Schematic of the pump–probe geometry. The resonantly excited A_1 phonon mode in $LiNbO_3$ is also shown with a polar component along the crystal *c* axis. **b**, Electro-optic sampling measurement of the CEP-stable pump pulses, which are 150 fs long, centred at 17.5 THz with 4 THz bandwidth. **c**, **d**, Time-resolved polarization rotation of the 800 nm probe (**c**) and changes in the second-harmonic intensity (**d**), for high (solid lines) and low (dashed lines) excitation fields. FFT, fast Fourier transform; MIR, mid-infrared.

[1]Max Planck Institute for the Structure and Dynamics of Matter, 22761 Hamburg, Germany. [2]Department of Physics, University of Oxford, Clarendon Laboratory, Oxford OX1 3PU, UK.

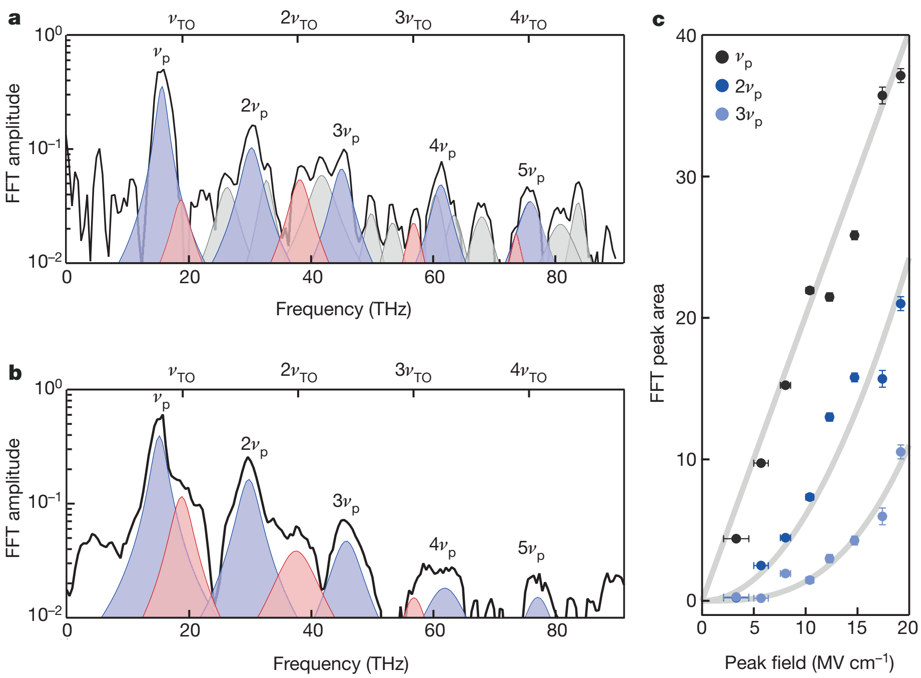


Figure 2 | Spectra of time-resolved optical responses and harmonic field dependences. **a**, **b**, FFT amplitude spectra of the polarization rotation (**a**) and second-harmonic intensity (**b**) measurements, for the high excitation field shown in Fig. 1. The blue peaks and red peaks correspond to multiples of the phonon-polariton frequencies ν_p (15.3 THz at 800 nm, 16.2 THz at 400 nm) and ν_{TO}, respectively. The grey peaks in **a** label sum and difference frequencies of ν_p and ν_{TO}, which are absent in the second-harmonic response. **c**, Excitation field dependence of the peak area at the first, second and third harmonic of ν_p, revealing a linear, quadratic and cubic dependence. Error bars represent the standard deviation σ calculated from the noise level of the experiment.

phase-matching between the probe light and the phonon-polariton propagating into the crystal[18,19].

At high pump fields (20 MV cm^{-1}, Fig. 1c, d, solid lines), a strongly anharmonic response was observed, with asymmetric oscillations in both polarization rotation and second-harmonic signals. The corresponding amplitude spectra are shown in Fig. 2a, b. In addition to the fundamental frequency components, several harmonics appeared. The most pronounced peaks were found at multiples of the 15 THz phonon-polariton mode, visible up to $n = 5$ (75 THz). Correspondingly, the amplitudes of the first three harmonics at $\nu = 15$ THz, 30 THz and 45 THz displayed a linear, quadratic and cubic dependence on the excitation field (see Fig. 2c). The polarization rotation spectrum also exhibited peaks at the sum and difference frequencies of these harmonics (see Extended Data Fig. 5 for detailed assignments of all peaks). These data are reminiscent of what has been extensively reported in the literature in the context of non-resonant terahertz and mid-infrared harmonic generation[20–25]. Here, however, the harmonics appear at multiples of the phonon-polariton frequency, instead of the central frequency of the optical pump field, indicating a different physical origin.

To analyse these data, we first consider the local lattice response. We start from the anharmonic lattice potential of the driven mode at ω_{TO} and ignore phonon-polariton propagation. *Ab initio* density functional theory (DFT) calculations (see Methods) yield the anharmonic lattice potential plotted in Fig. 3. This potential can be fitted by

$$U(Q_{IR}) = \frac{1}{2}\omega_{TO}^2 Q_{IR}^2 + \frac{1}{3}a_3 Q_{IR}^3 + \frac{1}{4}a_4 Q_{IR}^4 + \frac{1}{5}a_5 Q_{IR}^5 \qquad (1)$$

where Q_{IR} denotes the amplitude of the infrared-active mode, $\omega_{TO} = 2\pi\nu_{TO}$, and a_3, a_4 and a_5 are the coefficients of the cubic, quartic and quintic potential terms. Note that, in the potential of equation (1), we have omitted all terms that describe the coupling to other vibrational modes Q_j of the form $\sum_j Q_{IR}^2 Q_j$ (refs 5, 8, 9). These terms displace the average lattice structure along all the coupled coordinates Q_j and renormalize the eigenfrequency of the driven mode Q_{IR}. However, as shown in Extended Data Fig. 6, the effect is small and will not be discussed here.

Starting from the potential energy of equation (1), we derive the equation of motion for Q_{IR}, considering excitation with a mid-infrared light pulse of carrier frequency ω_{MIR} and duration T

$$\ddot{Q}_{IR} + 2\gamma\dot{Q}_{IR} + \omega_{TO}^2 Q_{IR} + a_3 Q_{IR}^2 + a_4 Q_{IR}^3 + a_5 Q_{IR}^4 = Z^* E(t) \qquad (2)$$

Here, Z^* denotes the effective charge of the phonon mode, γ is a dissipation constant, and $E(t) = E_0 \sin(\omega_{MIR} t)\exp(-t^2/T^2)$ is the excitation pulse profile. The calculated dynamics at the field strengths of 20 MV cm^{-1} used in the experiment are reported in Fig. 3 and predict peaks at harmonics of the fundamental frequency ν_{TO}.

A more comprehensive description of our experimental observations was obtained when propagation effects were taken into account. Finite-difference time-domain (FDTD)[26] simulations of phonon-polariton

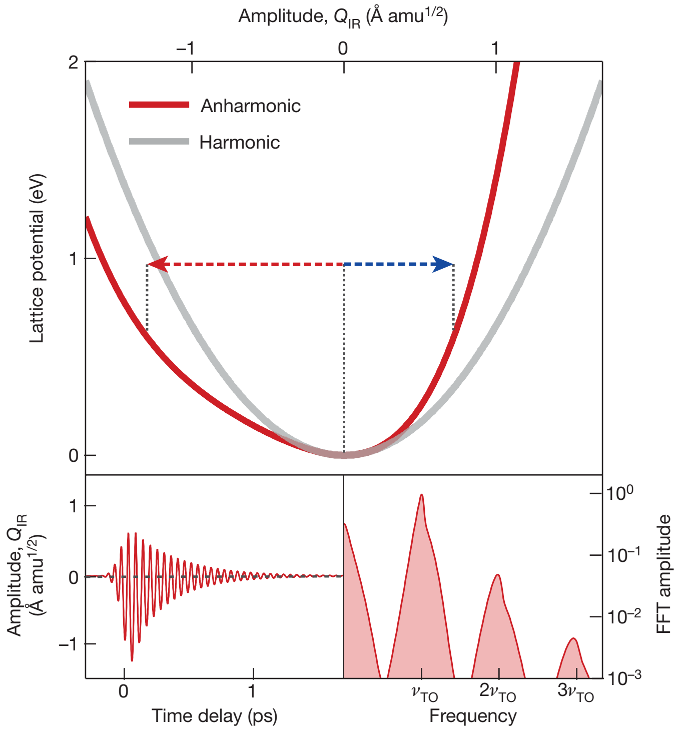


Figure 3 | Calculated A$_1$-mode potential energy. The calculated lattice potential energy of LiNbO$_3$ (red) of the A$_1$ mode shown in Fig. 1a is compared to a harmonic potential (grey) with the same fundamental frequency ω_{TO}. The arrows denote the expected positive and negative excursions for an energy of 0.6 eV, corresponding to the energy deposited per unit cell by the excitation pulses. The lower plots show the solution of the equation of motion and its amplitude spectrum.

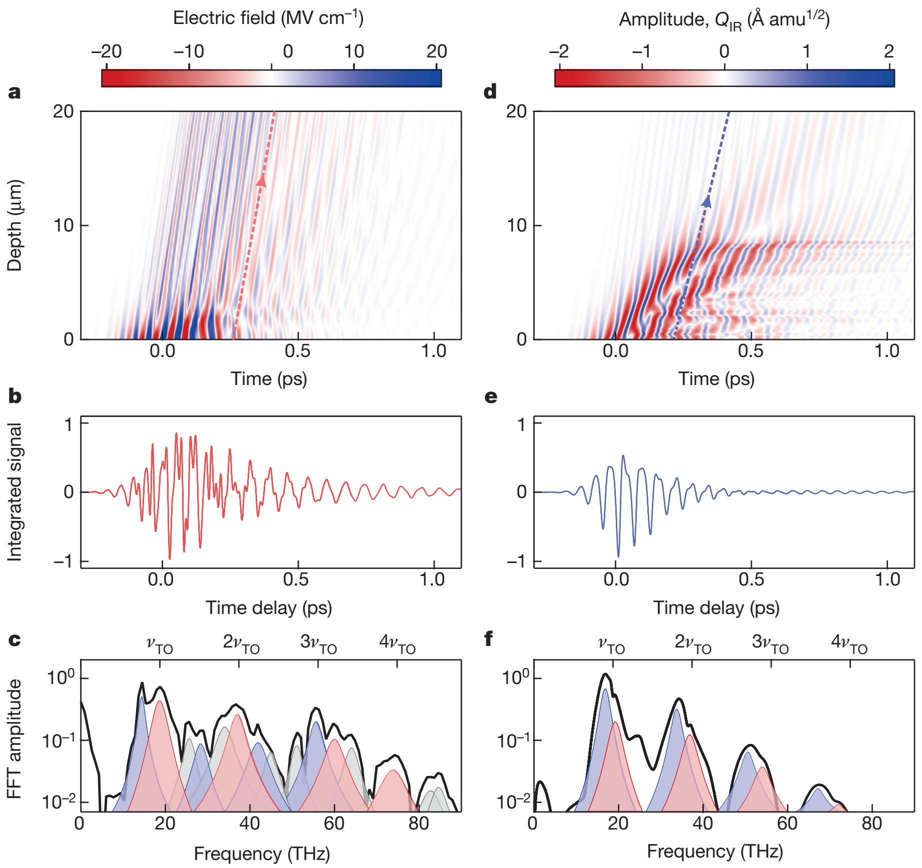


Figure 4 | FDTD simulations for the phonon-polariton propagation. a, Contour plot of the electric field as a function of depth d and time t inside LiNbO$_3$ after mid-infrared excitation. The red dashed line shows the propagation of the 800 nm probe pulse for one pump–probe time delay τ, following the relation $d_{800} = v_g t$. **b**, **c**, Time trace derived by integrating along the dashed red line in **a** for all pump–probe delays (**b**), and the corresponding amplitude spectrum (**c**). The spectrum shows harmonics of ν_p and ν_{TO} as well as mixed frequencies. **d**, Contour plot of the vibrational amplitude Q_{IR} after the same mid-infrared excitation as in **a**. The dashed blue line shows the propagation of the second-harmonic light at 400 nm, also for a single time delay τ. **e**, Time trace derived by spatially integrating the time derivative of Q_{IR} along the dashed blue line within the first 2 μm for all pump–probe delays. **f**, Amplitude spectrum of **e**, which shows broad peaks only at the harmonics of ν_p and ν_{TO}.

propagation are reported in Fig. 4. In these simulations, we combined the linear optical properties of LiNbO$_3$ (Extended Data Table 1) with the nonlinear lattice potential of equation (1) (see Methods and Extended Data Fig. 7). Figure 4a displays the amplitude of the propagating electric field as a function of sample depth d and time t. Both the phonon-polaritons and the broadband radiation emitted from the anharmonic motions propagate from the surface into the bulk, following the dispersion imposed by the material. By integrating the simulated electric field along the 800 nm light line, $d_{800} = v_g t$ (red dashed line in Fig. 4a, with v_g the group velocity), for each pump–probe time delay τ, we extracted the response shown in Fig. 4b, yielding good qualitative agreement with the polarization rotation measurement (compare Figs 4b and 1c). Figure 4c displays the corresponding amplitude spectrum, which comprises peaks at all sum and difference frequencies of the polariton and the transverse-optical mode, also in good agreement with experiment (see Fig. 2a).

We next turn to the key results of this paper, which are extracted from the time-dependent changes in the second-harmonic intensity $I_{SH}(\tau)$. As discussed elsewhere[12,13,27], a coherent phonon of frequency Ω generates frequency-shifted radiation in the second-harmonic field E_{SH} because of hyper-Raman scattering. Crucially, the detected spectral interferometry signal is proportional to the lattice velocity, $I_{SH}(\tau) = B\dot{Q}(\tau)$ (see also Methods and refs 27, 28). Therefore, we can compare the simulations of Fig. 4e, f with the experiments of Figs 1d and 2b by spatially integrating the time derivative of the simulated lattice coordinate $Q(t, z)$ along the 400 nm light line (Fig. 4d, dashed blue line). This integral was taken over the first 2 μm beneath the surface, where the second-harmonic light is generated in the experiment[8]. The corresponding simulated signal $I_{SH}(\tau)$ (Fig. 4e) contains frequency components at multiples of 16 THz and 19 THz (see Fig. 4f), in agreement with the measured data of Fig. 2b.

Most importantly, from the knowledge of $\dot{Q}(\tau)$, the microscopic lattice potential $U(Q)$ explored during each oscillation cycle could be reconstructed. We consider the coherent dynamics of the lattice at times after the pump pulse, that is, when no force is being applied onto the mode. For weak phonon damping ($\gamma \ll \Omega$), the total energy of the unforced oscillating lattice can be approximated as being constant over each cycle, $U(\tau) + E_{kin}(\tau) = \varepsilon$. Hence, we could retrieve the instantaneous potential energy $U(\tau) = \varepsilon - E_{kin}(\tau)$ from the knowledge of the kinetic energy, which is in turn proportional to the square of the measured second-harmonic signal $E_{kin}(\tau) = \frac{1}{2}\dot{Q}(\tau)^2 = \frac{1}{2}I_{SH}(\tau)^2/B^2$. The instantaneous potential energy $U(\tau)$, which was known except for a proportionality term $1/B^2$, could then be converted into $U(Q)$ by a time integral of the second-harmonic signal $I_{SH}(\tau) = B\dot{Q}(\tau)$, which yielded $Q(\tau)$. Hence, we could extract the shape of the lattice potential apart from a single proportionality constant. Because different cycles with different amplitudes and different total energy ε trace fractions of the potential energy $U(Q)$ many times, the shape of the potential reconstructed in this way was highly over-determined.

Figure 5 compares the lattice potential of the A$_1$ mode calculated from DFT (grey line) to the reconstructed potential (filled circles). The calculated and reconstructed curves were matched by adjusting one free parameter B (see Methods for details). Within the systematic uncertainties of DFT calculations (light grey shaded area), we find agreement between the shapes of the anharmonic potentials up to the highest amplitudes reached experimentally.

The tomography of the force field discussed above is straightforwardly extensible to all materials with a large bandgap, such as ferroelectrics, for which acceleration of quasiparticles in the field is neglected to first order. We note that direct measurements of the coordinate $Q(\tau)$ with femtosecond X-ray diffraction, for example from a free electron laser with pulses that are appropriately synchronized with the absolute phase of a strong terahertz field, would allow an unbiased measurement of the potential, without the need for comparing the data to a calculated potential and determining the constant. Also, full reconstruction of the force field of a material with N atoms

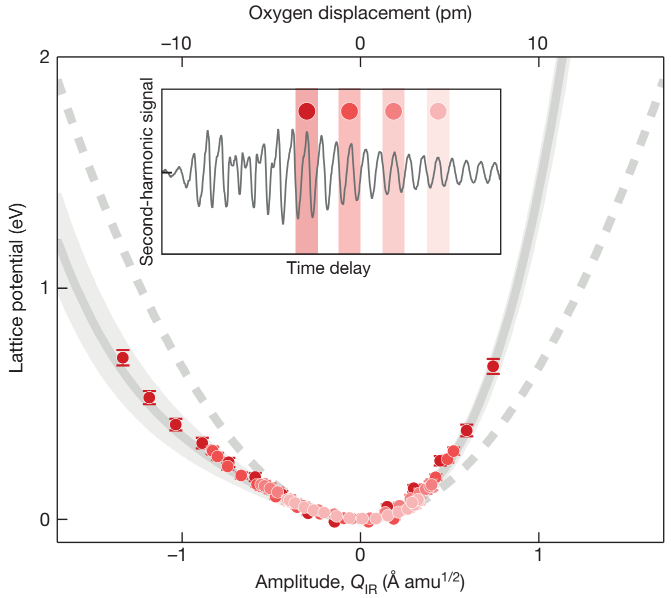


Figure 5 | Reconstructed A$_1$-mode potential energy. The potential energy of the A$_1$ mode (red filled circles) was reconstructed from different cycles of the time-resolved second-harmonic measurement shown in the inset. Error bars represent the standard deviation σ calculated from the noise level of the experiment. The grey solid line is the mode potential obtained by DFT calculations, and the light grey shaded area is an estimate of its systematic uncertainties. The experimental potential is scaled to the calculated potential using a single scaling factor (see Methods). From this comparison, we estimate maximum mode excursions of 1.4 Å amu$^{1/2}$ (where amu is atomic mass unit), corresponding to displacements of the oxygen atoms by about 14 picometres from their equilibrium positions. The dashed grey curve is the potential in the harmonic approximation.

requires the measurement of $3N-3$ lattice modes without symmetry considerations. Recent advances in the generation of mid-infrared and terahertz pulses that are both widely tunable and intense[29] make these prospects realistic. Tomographic measurements of force potentials in the vicinity of equilibrium phase transitions will yield crucial information not accessible otherwise. Finally, as the sampling of the potential can be retrieved within one cycle of the pump light, we envisage measurements of rapidly evolving potential energy surfaces.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 June; accepted 4 December 2017.
Published online 21 February 2018.

1. Shen, Y. R. *The Principles of Nonlinear Optics* (Wiley-Interscience, 2002).
2. Kampfrath, T., Tanaka, K. & Nelson, K. A. Resonant and nonresonant control over matter and light by intense terahertz transients. *Nat. Photon*. **7,** 680–690 (2013).
3. Nicoletti, D. & Cavalleri, A. Nonlinear light–matter interaction at terahertz frequencies. *Adv. Opt. Photonics* **8,** 401–464 (2016).
4. Mitrano, M. *et al.* Possible light-induced superconductivity in K_3C_{60} at high temperature. *Nature* **530,** 461–464 (2016).
5. Mankowsky, R. *et al.* Nonlinear lattice dynamics as a basis for enhanced superconductivity in $YBa_2Cu_3O_{6.5}$. *Nature* **516,** 71–73 (2014).
6. Nova, T. F. *et al.* An effective magnetic field from optically driven phonons. *Nat. Phys.* **13,** 132–136 (2017).
7. Fausti, D. *et al.* Light induced superconductivity in a stripe-ordered cuprate. *Science* **331,** 189–191 (2011).
8. Mankowsky, R., von Hoegen, A., Först, M. & Cavalleri, A. Ultrafast reversal of the ferroelectric polarization. *Phys. Rev. Lett.* **18,** 197601 (2017).
9. Först, M. *et al.* Nonlinear phononics as a new ultrafast route to lattice control. *Nat. Phys.* **7,** 854–856 (2011).
10. Schaufele, R. F. & Weber, M. J. Raman scattering by lithium niobate. *Phys. Rev.* **152,** 705–708 (1966).
11. Barker, A. S. & Loudon, R. Dielectric properties and optical phonons in $LiNbO_3$. *Phys. Rev.* **158,** 433–445 (1967).
12. Chang, Y.-M., Xu, L. & Tom, H. W. K. Observation of local-interfacial optical phonons at buried interfaces using time-resolved second-harmonic generation. *Phys. Rev. B* **59,** 12220–12223 (1999).
13. Denisov, V. N., Mavrin, B. N. & Podobedov, V. B. Hyper-Raman by vibrational excitations in crystals, glasses and liquids. *Phys. Rep.* **151,** 1–92 (1987).
14. Sell, A., Leitenstorfer, A. & Huber, R. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. *Opt. Lett.* **33,** 2767–2769 (2008).
15. Keiber, S. *et al.* Electro-optic sampling of near-infrared waveforms. *Nat. Photon.* **10,** 159–162 (2016).
16. Porer, M., Ménard, J.-M. & Huber, R. Shot noise reduced terahertz detection via spectrally postfiltered electro-optic sampling. *Opt. Lett.* **39,** 2435–2438 (2014).
17. Kojima, S., Kanehara, K., Hoshina, T. & Tsurumi, T. Optical phonons and polariton dispersions of congruent $LiNbO_3$ studied by far-infrared spectroscopic ellipsometry and Raman scattering. *Jpn. J. Appl. Phys.* **55,** 10TC02 (2016).
18. Claus, R. Light scattering by optical phonons and polaritons in perfect crystals. *Phys. Status Solidi B* **50,** 11–32 (1972).
19. Dastrup, B. S., Hall, J. R. & Johnson, J. A. Experimental determination of the interatomic potential in $LiNbO_3$ via ultrafast lattice control. *Appl. Phys. Lett.* **110,** 162901 (2017).
20. Ghimire, S. *et al.* Observation of high-order harmonic generation in a bulk crystal. *Nat. Phys.* **7,** 138–141 (2011).
21. Schubert, O. *et al.* Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nat. Photon.* **8,** 119–123 (2014).
22. Zaks, B., Liu, R. B. & Sherwin, M. S. Experimental observation of electron–hole recollisions. *Nature* **483,** 580–583 (2012).
23. Vampa, G. *et al.* Linking high harmonics from gases and solids. *Nature* **522,** 462–464 (2015).
24. Itatani, J. *et al.* Tomographic imaging of molecular orbitals. *Nature* **432,** 867–871 (2004).
25. Langer, F. *et al.* Lightwave-driven quasiparticle collisions on a subcycle timescale. *Nature* **533,** 225–229 (2016).
26. Taflov, A. & Hagness, S. C. *Computational Electrodynamics: The Finite-Difference Time-Domain Method* (Artech House, 2000).
27. Yan, Y. X., Gamble, E. B. Jr & Nelson, K. A. Impulsive stimulated scattering: General importance in femtosecond laser pulse interactions with matter, and spectroscopic applications. *J. Chem. Phys.* **83,** 5391–5399 (1985).
28. Merlin, R. Generating coherent THz phonons with light pulses. *Solid State Commun.* **102,** 207–220 (1997).
29. Liu, B. *et al.* Generation of narrowband, high-intensity, carrier-envelope phase stable pulses tunable between 4 and 18 THz. *Opt. Lett.* **42,** 129–131 (2017).

Acknowledgements We thank R. Merlin and M. Altarelli for discussions. The research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement No. 319286 (QMAC). We acknowledge support from the Deutsche Forschungsgemeinschaft via the excellence cluster 'The Hamburg Centre for Ultrafast Imaging—Structure, Dynamics and Control of Matter at the Atomic Scale' and the Priority Program SFB925 'Light induced Dynamics and Control of Correlated Quantum Systems'.

Author Contributions A.C., together with A.v.H. and R.M., conceived this project. R.M., A.v.H. and M. Först built the experimental set-up. A.v.H. and R.M. conducted the experiment and analysed the data. M. Fechner performed the DFT calculations. A.v.H. conducted the FDTD simulation. All authors interpreted the data and contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.A. (andrea.cavalleri@mpsd.mpg.de) or A.v.H. (alexander.von-hoegen@mpsd.mpg.de).

Reviewer Information *Nature* thanks M. Kira and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Experimental set-up. The CEP-stable, 150-fs-long, 17.5 THz mid-infrared pump pulses of 4 THz bandwidth were obtained by mixing the two signal beams from two optical parametric amplifiers, which were seeded by the same white light and pumped by 30 fs, 800 nm pulses at a repetition rate of 1 kHz. The nonlinear lattice dynamics in $LiNbO_3$ were probed by a time-delayed replica of the 800 nm pulses, in non-collinear geometry with an angle of 30° to the mid-infrared pump (see Extended Data Fig. 1). The CEP stability of the excitation pulse is reflected in phase stability of the resonantly driven coherent oscillations of the A_1 phonon mode.

The pump-induced polarization rotation (PR) of the 800 nm beam was measured by detecting the time-resolved difference signal of two intensity-balanced photodiodes placed behind a half-wave plate and a Wollaston prism.

Owing to the large second-order nonlinear susceptibility of $LiNbO_3$, the 800 nm probe pulses also generated second-harmonic (SH) light at 400 nm, which was separated from the fundamental beam after the sample by a dichroic mirror and detected with a photomultiplier tube. The SH signal originates from a layer of one coherence length $l_c = 1.3\,\mu$m below the surface[8,30].

All experiments were conducted at room temperature. The sample used in the experiments was a commercially available congruent $LiNbO_3$ single crystal (5 mm × 5 mm × 5 mm).

PR and SH detection processes. The nonlinear interaction of a lattice vibrational mode with an optical probe pulse involves Raman scattering in PR measurements and hyper-Raman scattering in SH measurements[12,13,27,28]. These processes can be described by the wave equation

$$\frac{\partial^2 E}{\partial z^2} - \frac{n^2}{c^2}\frac{\partial^2 E}{\partial t^2} = \rho \frac{\partial^2 (E_{\mathrm{probe}} Q)}{\partial t^2}$$

where n is the refractive index of the material and c the vacuum speed of light. The constant ρ contains the number density of oscillators and the Raman cross-section. E_{probe} denotes the probe laser field at angular frequency ω_{probe} (refs 27, 28), that is, the fundamental 800 nm light in the PR measurement and the 400 nm light in the SH measurement. $Q(z, t+\tau) = Q_0(z)\sin[\Omega(t+\tau)]$ is the time-dependent amplitude of the excited vibrational mode, with τ the time delay between pump and probe pulses. A general solution to this equation is[27,28]

$$E(z,t) = E_{\mathrm{probe}}(z,t) - a\frac{\partial}{\partial t}(Q(z,t+\tau)E_{\mathrm{probe}}(z,t))$$

showing that the probe electric field is modulated by the time derivative of $Q(z, t+\tau)E_{\mathrm{probe}}(z, t)$. After interaction with a phonon-polariton of frequency Ω and at a specific time delay, the spectrum of the transmitted probe, which is the Fourier transform of $E(z, t)$, reads

$$E(\omega) = E_{\mathrm{probe}}(\omega) + \beta\omega[E_{\mathrm{probe}}(\omega+\Omega)\exp(i\Omega\tau) - E_{\mathrm{probe}}(\omega-\Omega)\exp(-i\Omega\tau)]$$

It contains the unperturbed probe spectrum $E_{\mathrm{probe}}(\omega)$ and sidebands generated at $\omega_{\mathrm{probe}} \pm \Omega$. Importantly, these sidebands acquire a time-delay-dependent phase $\exp(\pm i\Omega\tau)$. Their phase-sensitive detection, for example achieved by spectral interference with the local oscillator $E_{\mathrm{probe}}(\omega)$ on the detector, carries information about both phase and amplitude of the phonon-polariton[15,16].

In this case, the measured intensity at a time delay τ is

$$\begin{aligned} I(\tau) = \int \mathrm{d}\omega\{|E_{\mathrm{probe}}(\omega) + \beta\omega[E_{\mathrm{probe}}(\omega+\Omega)\exp(i\Omega\tau) \\ -E_{\mathrm{probe}}(\omega-\Omega)\exp(-i\Omega\tau)]|^2\} \\ = I_{\mathrm{probe}} + \alpha\Omega\cos(\Omega\tau) + \gamma\Omega^2\cos(2\Omega\tau) \end{aligned}$$

We disregard the homodyne component proportional to $\gamma\Omega^2$, which is far smaller than the heterodyne component, proportional to $\alpha\Omega$ (ref. 32). Hence, the detected interference signal is[27,28,31]

$$I(\tau) = I_{\mathrm{probe}} + \alpha\Omega\cos(\Omega\tau)$$

Crucially, this time-delay-dependent signal is phase-shifted by $\pi/2$ and amplitude scaled proportional to Ω with respect to the lattice vibration $Q(t) = Q_0\sin[\Omega(t+\tau)]$. Hence, it is proportional to the velocity of the vibrational motion $\dot{Q}(\tau)$. For in-depth discussion, we refer to refs 27, 28 and 31.

In the PR measurements, the detected difference signal is $\Delta I(\tau) = I_{\parallel}(\tau) - I_{\perp}(\tau) = 2\alpha\Omega\cos(\Omega\tau)$

The time resolution is determined by the bandwidth of the local oscillator at the detector[15,16] which is spectrally broadened with respect to the incident pulse owing to self-phase modulation in the $LiNbO_3$ crystal[32,33] (see Extended Data Fig. 3a). The sampling efficiency, calculated according to ref. 15 and shown in Extended Data Fig. 3b, allows efficient detection up to 80 THz. Further, the interaction length between the 800 nm probe and the phonon-polariton harmonics is determined by the penetration depths, which increase with increasing harmonic order (see also Extended Data Fig. 3b). As result, the PR measurements cannot be used to quantify the amplitude of the atomic motions in a straightforward manner.

In the SH measurement, the detected light is generated in a thin layer extending to 1.3 μm below the sample surface[8,30]. Therefore, the interaction length with the phonon-polariton harmonics does not change for different harmonic orders. The SH bandwidth supports efficient detection up to 60 THz (see Extended Data Fig. 3c, d). A bandpass filter was used to shape the spectral response function in order to flatten the sampling efficiency for the first three harmonics[15,31].

The phonon-polariton induced oscillatory signal components were extracted from the $I_{\mathrm{SH}}(\tau)$ data via subtraction of a slowly varying background, which results from the modification of $\chi^{(2)}$ due to changes in the ferroelectric polarization (see ref. 8).

Phase-matching between probe light and phonon-polariton. The amplitude spectra shown in Extended Data Fig. 4 are well understood by considering the phase-matching between the probe light and the phonon-polariton propagating into the crystal. The phonon-polariton dispersion of $LiNbO_3$ is plotted as $\nu_{\mathrm{P}} = \frac{c_0}{\sqrt{\varepsilon(\nu)}}q$, where c_0 is the vacuum speed of light and $\varepsilon(\nu)$ the dielectric function. The light lines $\nu = v_{\mathrm{g}}q$ of the 800 nm ($v_{\mathrm{g},800} = c_0/2.3$) and the 400 nm ($v_{\mathrm{g},400} = c_0/3.03$) probe fields are also shown, where v_{g} and q denote the group velocity and wave number, respectively. Phase-matching occurs at those frequencies for which the light lines intersect the phonon-polariton dispersion curve[18,34], that is, at 15 THz (PR), 16 THz (SH) and 19 THz (both PR and SH).

Peak assignments in the PR amplitude spectrum. Extended Data Fig. 5 displays a detailed assignment of all peaks in the amplitude spectrum of the PR measurement. Blue and red colours indicate up and down shifts, corresponding to sum and difference frequency mixing, respectively.

Linear optical properties of $LiNbO_3$. The low-frequency linear optical properties for light polarized along the $LiNbO_3$ c axis are dominated by two optical phonon modes at 7.8 THz and 18.9 THz. They also include a weak mode at 8.2 THz and a feature at 21 THz which has been attributed to two-phonon absorption[11]. Extended Data Fig. 7 shows the terahertz reflectivity spectrum of the investigated sample, measured via Fourier transform infrared spectroscopy (FTIR), together with fits of four and two Lorentzian oscillators. The fit parameters for the two dominating optical phonons (listed in Extended Data Table 1) were used in the FDTD simulations of the phonon-polariton propagation. The reflectivity spectrum simulated from the parameters of these two oscillators agrees with the experimental data within the region of interest (12–20 THz).

FDTD phonon-polariton simulations. The phonon-polariton propagation dynamics in $LiNbO_3$ have been calculated by solving Maxwell's equations in space and time. To this end, we used FDTD in one spatial dimension[26].

We modelled the linear response of the material using the parameters of the two dominant optical phonons obtained from fitting the FTIR measurement (see above). For each mode, the equation of motion is given by

$$\ddot{Q}_{\mathrm{IR}} + 2\gamma\dot{Q}_{\mathrm{IR}} + \omega_{\mathrm{TO}}^2 Q_{\mathrm{IR}} = Z^* E(t)$$

Here, γ is the damping constant, ω_{TO} the phonon angular frequency and Z^* the phonon-mode effective charge, which can be expressed as $\omega_{\mathrm{TO}}\sqrt{\varepsilon_0 - \varepsilon_\infty}\sqrt{\epsilon_0/n}$ with n the oscillator density, ϵ_0 the vacuum permittivity, and ε_0 and ε_∞ the low-frequency and high-frequency limits of the dielectric function, respectively. The oscillator density was approximated as one oscillator per unit cell. For each mode, ε_0 and ε_∞ were derived from the generalized Lyddane–Sachs–Teller relation[35].

The above equation was solved at every discrete point of the grid in space and time using the values of the electric field calculated from Maxwell's equation. The oscillator equation and Maxwell's equation are coupled via the electric displacement field

$$D = \epsilon_0\varepsilon_\infty E + \omega_{\mathrm{TO}}\sqrt{\varepsilon_0 - \varepsilon_\infty}\sqrt{\epsilon_0 n}\; Q_{\mathrm{IR}}$$

The linear optical properties of $LiNbO_3$ are well reproduced by our simulation (see Extended Data Fig. 7).

Nonlinear effects were captured by introducing the lattice anharmonicities of the driven A_1 mode into the above equation of motion:

$$\ddot{Q}_{\mathrm{IR}} + 2\gamma\dot{Q}_{\mathrm{IR}} + \omega_{\mathrm{TO}}^2 Q_{\mathrm{IR}} + a_3 Q_{\mathrm{IR}}^2 + a_4 Q_{\mathrm{IR}}^3 + a_5 Q_{\mathrm{IR}}^4 = Z^* E(t)$$

The anharmonic coefficients a_3, a_4 and a_5 are taken from *ab initio* DFT calculations as described below ($a_3 = 1{,}567.65$ meV amu$^{-3/2}$ Å^{-3}, $a_4 = 900.8$ meV amu^{-2} Å^{-4}, $a_5 = 7.1$ meV amu$^{-5/2}$ Å^{-5}). Here, the mid-infrared pump pulse was set to a field strength of 30 MV cm^{-1}, carrier frequency 17.5 THz and duration 180 fs, comparable to the experiment.

We evaluated the equations in time steps of 0.5 fs and with a spatial grid of 0.5 μm. Perfectly matched boundary conditions were implemented to impede back reflection.

DFT calculations for the potential energy along the A$_1$ coordinate. To explore the nonlinear response of a resonantly excited phonon mode we performed first-principle computations within the framework of DFT. All our computations were carried out using DFT as implemented in the QUANTUM ESPRESSO code[36]. We used ultrasoft pseudopotentials, which contain as valence states the $2p2s$ for lithium, $4s^2 4p^6 4d^4 5s^1$ for niobium and $2s^2 2p^4$ for oxygen. As numerical parameters, we applied a cut-off energy of 80 Rydberg (Ry) for the plane-wave expansion and 400 Ry for the charge density. For all computations, we sampled the Brillouin zone with a 17 × 17 × 17 ***k***-point mesh generated with the Monkhorst and Pack scheme[37] and reiterated total energy calculations until the total energy became less than 10^{-10} Ry. Before calculating phonon-modes, we fully structurally relaxed the unit cell regarding forces and pressure below the threshold of 5 μRy per a_0. We performed density functional perturbation theory[38] calculations to obtain phonon-mode eigenvectors and frequencies of the phonon modes. Finally, we computed the anharmonic phonon potential by calculating the total energy for structures, which have been modulated with the phonon eigenvector. Least-mean-square fits of this total energy landscape reveal the anharmonic coefficients of equation (2) of the main text and the phonon-mode eigenvector as shown in Fig. 1a.

Scaling the reconstructed potential. The unknown proportionality factor B, which connects the measured SH signal to the vibrational velocity via $I_{\mathrm{SH}}(\tau) = B\dot{Q}(\tau)$, leaves a single scaling factor to the reconstruction. The kinetic energy becomes $E_{\mathrm{kin}}(\tau) = [I_{\mathrm{SH}}(\tau)/B]^2/2$ and the vibrational amplitude $Q(\tau) = \int [I_{\mathrm{SH}}(\tau)/B]\mathrm{d}\tau$. Hence, the y axis of the reconstruction will be scaled with B^2 and the x axis with B to the correct absolute values. This constant B can be derived by fitting the function $f(Q) = (1/B^2)U(BQ)$ to the experimental data, where $U(Q)$ is the potential obtained by DFT. Once B is retrieved, the experimental x axis and y axis can be rescaled to the absolute phonon amplitude in terms of Å amu$^{1/2}$ and the potential energy in eV, respectively.

The maximum displacement of the oxygen atoms involved in the A$_1$ vibrational mode was calculated with the knowledge of the phonon eigenvectors, which we obtained from DFT calculations. We find a maximum displacement of the oxygen atoms of approximately 14 picometres, which amounts to 7% of the Nb–O and 5% of the O–O nearest-neighbour distance at the corresponding potential energy (0.7 eV), which agrees with the estimated energy deposited per unit cell (0.6 eV at a pulse energy of 3 μJ).

Effects of $g_j Q_{\mathrm{IR}}^2 Q_j$ nonlinear phonon coupling. As well as the anharmonicity of the driven lattice mode, the full lattice potential also comprises nonlinear coupling to other phonon modes of the form $g_j Q_{\mathrm{IR}}^2 Q_j$:

$$U(Q_{\mathrm{IR}}, Q_j) = \frac{1}{2}\omega_{\mathrm{TO}}^2 Q_{\mathrm{IR}}^2 + \frac{1}{3}a_3 Q_{\mathrm{IR}}^3 + \frac{1}{4}a_4 Q_{\mathrm{IR}}^4 + \frac{1}{5}a_5 Q_{\mathrm{IR}}^5 + \sum_j \frac{1}{2}\omega_j^2 Q_j^2 + \sum_j g_j Q_{\mathrm{IR}}^2 Q_j$$

Here, Q_j denotes the amplitude of a coupled lattice mode and ω_j its resonance frequency[5,8,9]. For strongly driven Q_{IR}, the nonlinear interaction leads to a directional force on the coupled mode Q_j, which can be used to control the functionality of materials[8].

In addition, the finite amplitude Q_j renormalizes the fundamental frequency of Q_{IR}, as can be seen in the equations of motion:

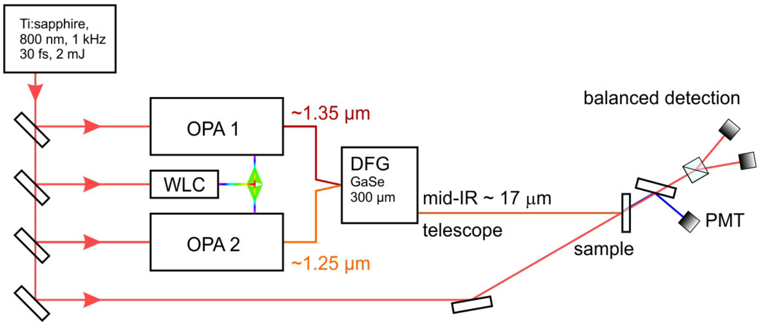
$$\ddot{Q}_{\mathrm{IR}} + 2\gamma\dot{Q}_{\mathrm{IR}} + (\omega_{\mathrm{TO}}^2 - 2gQ_j)Q_{\mathrm{IR}} = Z^* E(t)$$

$$\ddot{Q}_j + 2\gamma_j\dot{Q}_j + \omega_j^2 Q_j = g_j Q_{\mathrm{IR}}^2$$

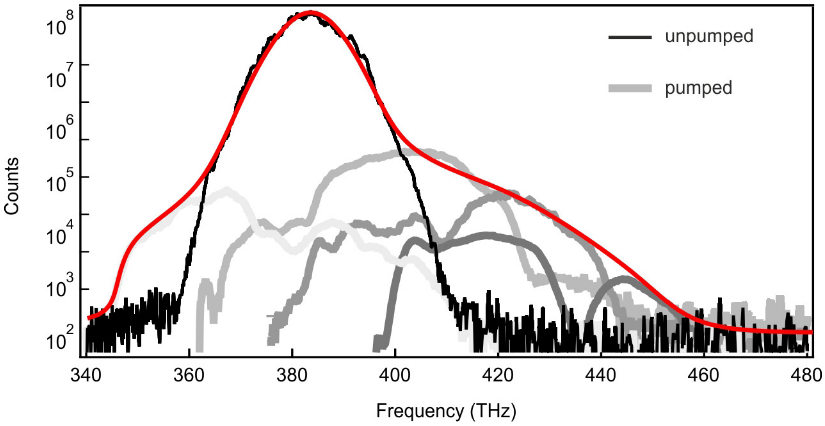
This frequency renormalization $\omega'_{\mathrm{TO}} = \sqrt{(\omega_{\mathrm{TO}}^2 - 2gQ_j)}$ was observed in our experiment with a maximum change of 3.5% at the highest driving field (see Extended Data Fig. 6).

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

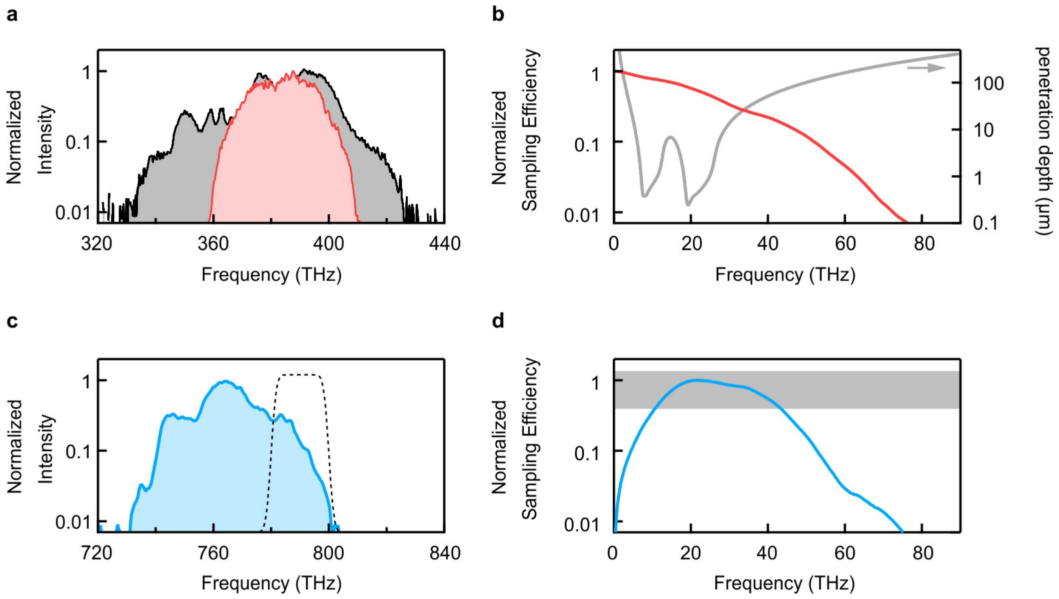
30. Mlejnek, M., Wright, E. M., Moloney, J. V. & Bloembergen, N. Second harmonic generation of femtosecond pulses at the boundary of a nonlinear dielectric. *Phys. Rev. Lett.* **83,** 2934 (1999).
31. Wahlstrand, J. K., Merlin, R., Li, X., Cundiff, S. T. & Martinez, O. E. Impulsive stimulated Raman scattering: comparison between phase-sensitive and spectrally filtered techniques. *Opt. Lett.* **30,** 926 (2005).
32. Träger, F. *Springer Handbook of Lasers and Optics* (Springer, 2007).
33. Dharmadhikari, J. A., Dota, K., Mathur, D. & Dharmadhikari, A. K. Spectral broadening in lithium niobate in a self-diffraction geometry using ultrashort pulses. *Appl. Phys. B* **140,** 122–129 (2016).
34. Stevens, T., Wahlstrand, J. K. & Merlin, R. Cherenkov radiation at speeds below the light threshold: phonon-assisted phase matching. *Science* **291,** 627–630 (2001).
35. Chaves, A. S. & Porto, S. P. S. Generalized Lyddane–Sachs–Teller relation. *Solid State Commun.* **13,** 865–868 (1973).
36. Giannozzi, P. *et al.* QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21,** 395502 (2009).
37. Monkhorst, H. & Pack, J. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13,** 5188 (1976).
38. Baroni, S., de Gironcoli, S., Dal Corso, A. & Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73,** 515 (2001).



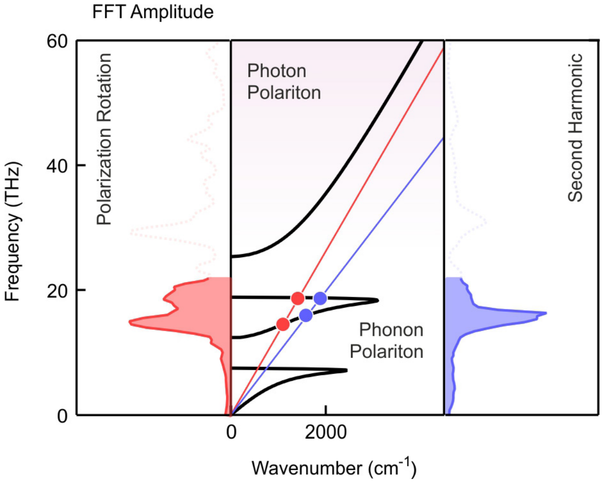
Extended Data Figure 1 | Experimental set-up. Pulses (30 fs) from a Ti:sapphire amplifier are used to pump two optical parametric amplifiers (OPA), which are seeded by the same white-light continuum (WLC). CEP-stable, 3 μJ, 150 fs pulses at 17 μm wavelength are obtained by difference frequency generation (DFG) of the two signal beams from the OPAs. The mid-infrared light is focused to a spot size of approximately 65 μm using a telescope and overlapped with the 800 nm probe beam (40 nJ, 35 μm spot size).



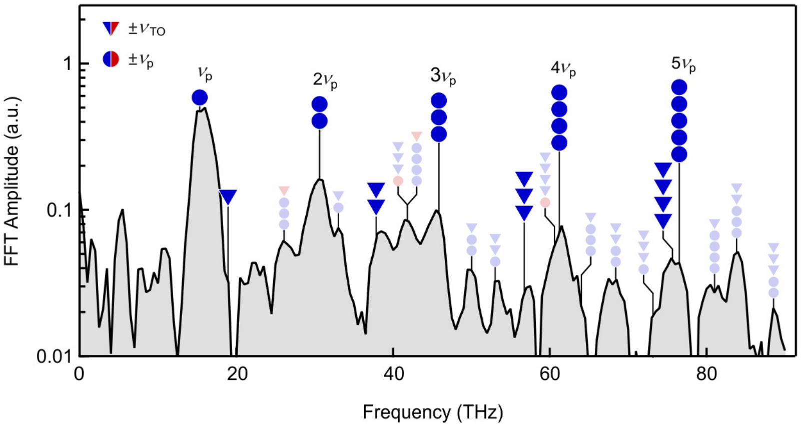
Extended Data Figure 2 | Sideband generation from phonon harmonics. The black solid line is the incident spectrum of the 800 nm probe pulses with a bandwidth of about 30 THz. The grey solid lines are the sidebands generated from the phonon harmonics measured at different positions behind the $LiNbO_3$ crystal. Owing to momentum conservation, each sideband propagates in a slightly different direction compared with the unperturbed 800 nm beam. The red line is a guide to the eye of the resulting spectral broadening.



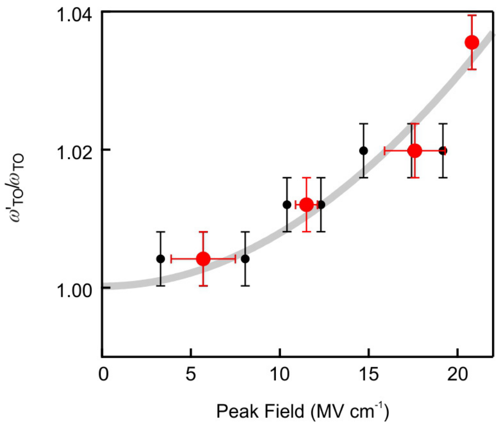
Extended Data Figure 3 | Probe spectra and sampling efficiencies. **a**, Spectrum of the 800 nm probe pulse before (red) and after (grey) propagation through the unpumped $LiNbO_3$ crystal in units of THz. **b**, Red curve: sampling efficiency of the 800 nm light calculated with the spectrum shown in **a**. The grey curve is the penetration depth in the mid-infrared region obtained from FTIR spectroscopy. **c**, Spectrum of the generated SH light (blue curve) and normalized transmission of the bandpass filter placed in front of the detector (dashed curve), also shown in units of THz. **d**, Sampling efficiency of the SH light with the spectrum shown in **c**. The sampling efficiency is almost constant in the 15–45 THz region of the first three phonon harmonics.



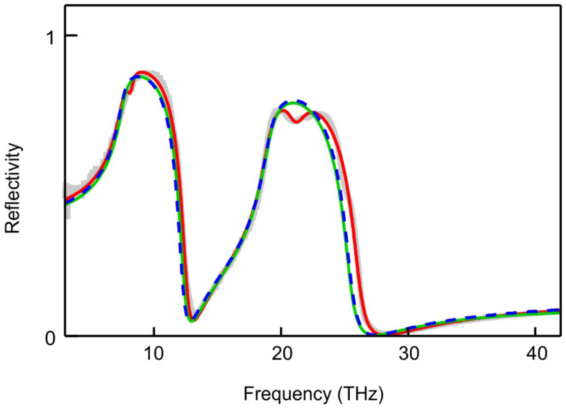
Extended Data Figure 4 | Phonon-polariton dispersion. The phonon-polariton dispersion of the two dominant lattice modes in $\mathrm{LiNbO_3}$ (black curve) and two light lines $\nu = v_g q$ for 800 nm (red) and 400 nm (blue) wavelengths are shown. The dots mark the points of intersection with the dispersion relation, which correspond to the observed fundamental frequencies of the driven mode (left and right panels).



Extended Data Figure 5 | Assignment of phonon harmonics. The amplitude spectrum of the time-resolved PR measurement is shown. Blue symbols denote a blueshift of 15 THz (triangles) and 19 THz (circles). Multiple symbols represent shifts by multiples of the corresponding frequencies. Red symbols denote redshifts.



Extended Data Figure 6 | Phonon frequency renormalization. The black circles denote the peak-field-dependent fundamental phonon frequencies extracted from Fourier transformations of the time-resolved signals. Values at the same frequency have been binned (red circles) to account for the limited frequency resolution of the FFT analysis. The error bars denote 1σ (67% confidence interval). The grey line is a fit to the data with the function $f(E) = \sqrt{1 + aE^2}$.



Extended Data Figure 7 | Terahertz reflectivity spectrum. The grey solid line is the measured terahertz reflectivity spectrum of $LiNbO_3$ with light polarized along the *c* axis. The red line is a fit considering four Lorentzian oscillators. The dashed blue line is a fit considering only the two dominant phonon modes at 7.5 THz and 19 THz. The green line is the FDTD simulated reflectivity when only these two oscillators are considered (see Methods).

Extended Data Table 1 | Parameters for the A_1 Lorentzian oscillator

Oscillator #	Frequency (cm^{-1})	Oscillator strength (cm^{-1}) $\omega_{TO}\sqrt{\varepsilon_0-\varepsilon_\infty}$	Damping (cm^{-1})
1	249.3	922.8	27.7
2	271.6	384.1	20
3	632	955.9	33.5
4	696.7	352.5	76.2
ε_∞	4.4054		

Frequencies, oscillator strengths and damping constants of the $LiNbO_3$ vibrational modes were obtained from a fit of four Lorentzian oscillators to the reflectivity spectrum.

doi:10.1038/nature25494

Skin electronics from scalable fabrication of an intrinsically stretchable transistor array

Sihong Wang[1]*, Jie Xu[1]*, Weichen Wang[2], Ging-Ji Nathan Wang[1], Reza Rastak[3], Francisco Molina-Lopez[1], Jong Won Chung[1,4], Simiao Niu[1], Vivian R. Feig[2], Jeffery Lopez[1], Ting Lei[1], Soon-Ki Kwon[5], Yeongin Kim[6], Amir M. Foudeh[1], Anatol Ehrlich[1], Andrea Gasperini[1], Youngjun Yun[1,4], Boris Murmann[6], Jeffery B.-H. Tok[1] & Zhenan Bao[1]

Skin-like electronics that can adhere seamlessly to human skin or within the body are highly desirable for applications such as health monitoring[1,2], medical treatment[3,4], medical implants[5] and biological studies[6,7], and for technologies that include human–machine interfaces, soft robotics and augmented reality[8,9]. Rendering such electronics soft and stretchable—like human skin—would make them more comfortable to wear, and, through increased contact area, would greatly enhance the fidelity of signals acquired from the skin. Structural engineering of rigid inorganic and organic devices has enabled circuit-level stretchability, but this requires sophisticated fabrication techniques and usually suffers from reduced densities of devices within an array[2,10–12]. We reasoned that the desired parameters, such as higher mechanical deformability and robustness, improved skin compatibility and higher device density, could be provided by using intrinsically stretchable polymer materials instead. However, the production of intrinsically stretchable materials and devices is still largely in its infancy[13–15]: such materials have been reported[11,16–19], but functional, intrinsically stretchable electronics have yet to be demonstrated owing to the lack of a scalable fabrication technology. Here we describe a fabrication process that enables high yield and uniformity from a variety of intrinsically stretchable electronic polymers. We demonstrate an intrinsically stretchable polymer transistor array with an unprecedented device density of 347 transistors per square centimetre. The transistors have an average charge-carrier mobility comparable to that of amorphous silicon, varying only slightly (within one order of magnitude) when subjected to 100 per cent strain for 1,000 cycles, without current–voltage hysteresis. Our transistor arrays thus constitute intrinsically stretchable skin electronics, and include an active matrix for sensory arrays, as well as analogue and digital circuit elements. Our process offers a general platform for incorporating other intrinsically stretchable polymer materials, enabling the fabrication of next-generation stretchable skin electronic devices.

Electronics on human skin generally comprise, but are not limited to, two types of component: input/output devices for human interaction (the input might be, for example, a sensor element; the output might be a display), and electronic circuits for information processing. So far, stretchability has been demonstrated for certain input/output devices[20–24], but there are still no functional skin-like stretchable circuits, mostly because of the much higher complexity required at both circuit and device levels. Therefore, realizing skin electronics will rely on the development of intrinsically stretchable circuits composed of densely integrated transistors.

Fabricating intrinsically stretchable electronics (Fig. 1a) requires intrinsically stretchable materials, especially stretchable semiconductors and conductors that possess uncompromised electrical conduction, even under large strains. Recently, polymers have been shown to be the most promising materials family for enabling both high electrical performance and intrinsic stretchability[11,16–19]. Moving forwards, a fabrication technology is needed that incorporates these materials into mass-producible arrays of transistors. However, polymer-based intrinsically stretchable electronic materials are solution deposited, and are highly susceptible to damage from both organic solvents and ultraviolet light. In general, they are incompatible with standard photolithography microfabrication technology, making it highly challenging to produce functional electronics *en masse* beyond individual transistors. Another obstacle lies in the incorporation of new materials, which typically necessitates an entirely new fabrication process. Therefore, a universal fabrication process or platform for generating intrinsically stretchable transistor arrays is needed to move materials development systematically to electronics and, finally, to desired applications. Here, we describe a fabrication platform with high yield and uniformity, which results in intrinsically stretchable transistor arrays (Fig. 1a) with a density and device count that greatly surpass those achieved by strain engineering and other approaches (Extended Data Table 1). Highlighting its versatility, we use this platform to demonstrate various intrinsically stretchable electronics ranging from active matrices to analogue and digital circuits. Further integrating these circuit components together with stretchable input/output devices could finally bring about skin electronic systems.

With its stretchability enabled by intrinsically stretchable materials—rather than through geometry design with rigid materials—our fabricated stretchable transistor array (Fig. 1b) has a record transistor density of about 347 transistors per square centimetre, and has high stretchability (up to 100% while maintaining a charge-carrier mobility of $0.98\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$). Moreover, the fabrication process can be scaled to produce a large array of 6,300 transistors over an area of around $4.4 \times 4.4\,\mathrm{cm}^2$ (Fig. 1c); this array is semi-transparent and has skin-like conformability and stretchability. This transistor array platform has enabled the first realization of skin-like intrinsically stretchable circuits (Fig. 1d), as the basic components of skin electronics.

The general fabrication process flow is designed to enable high device yield, device-to-device uniformity, good material compatibility between layers, and good electrical and mechanical performances. Specifically, the process flow (Fig. 2a) uses layer-by-layer direct deposition of active components to avoid the low yield and poor uniformity typically obtained from transfer processes that rely on peeling off through physical adhesion. First, to ensure good substrate compatibility during all the fabrication processes, we begin with a silicon wafer that is coated with a water-soluble sacrificial layer (dextran) to enable final release of the devices onto a stretchable substrate. Next, a stretchable dielectric is deposited and photo-patterned. Then, a stretchable

[1]Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA. [2]Department of Materials Science and Engineering, Stanford University, Stanford, California 94305, USA. [3]Department of Civil and Environmental Engineering, Stanford University, Stanford, California 94305, USA. [4]Samsung Advanced Institute of Technology, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-803, Republic of Korea. [5]Department of Materials Engineering and Convergence Technology and ERI, Gyeongsang National University, Jinju, 660-701, Republic of Korea. [6]Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA.
*These authors contributed equally to this work.

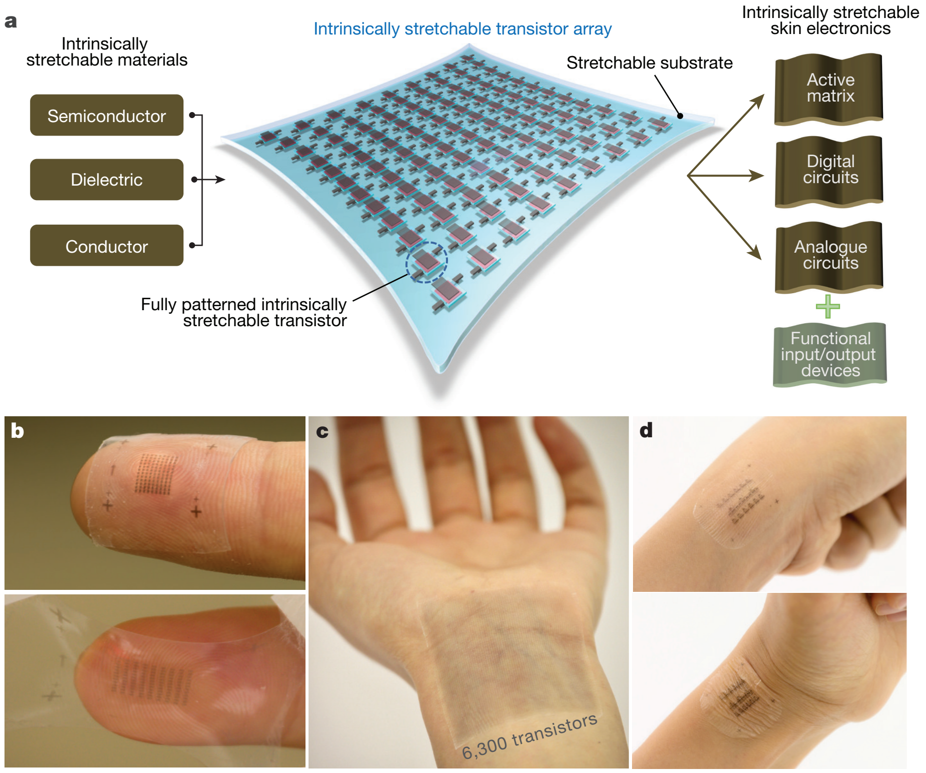

a
Intrinsically stretchable materials
Semiconductor
Dielectric
Conductor
Intrinsically stretchable transistor array
Stretchable substrate
Fully patterned intrinsically stretchable transistor
Intrinsically stretchable skin electronics
Active matrix
Digital circuits
Analogue circuits
Functional input/output devices
b
c
6,300 transistors
d

Figure 1 | Intrinsically stretchable transistor array as a core platform for functional skin electronics. a, Three-dimensional diagram of an intrinsically stretchable transistor array as the core building block of skin electronics. **b**, An array of 108 stretchable transistors on a fingertip, showing an unprecedented device density of 347 transistors per cm^2. **c**, A large-scale array of intrinsically stretchable transistors—with 6,300 transistors in an area of around 4.4 × 4.4 cm^2—attached conformably to a human inner wrist. **d**, Intrinsically stretchable circuits—a prototype for highly conformable functional skin electronics—attached to a bent human wrist.

semiconductor and a stretchable conductor for source/drain electrodes (together with the first layer of interconnect) are deposited and patterned consecutively, to get a top-contact structure. A stretchable substrate is laminated onto the device, which is then soaked in water to release it from the rigid substrate. Finally, the gate electrodes (and the second layer of interconnect) are deposited and patterned on the dielectric layer to complete the transistor structure.

Below we discuss the key fabrication steps in greater detail. First, the photo-patterned dielectric layer (Extended Data Fig. 1) needs to be solvent resistant to allow semiconductor deposition. To realize this with general applicability to stretchable dielectrics, we took advantage of azide-crosslinking chemistry[25,26], which can be initiated by ultraviolet light, and which is based on a reaction between azide groups and the CH group that is commonly found in elastomers (Fig. 2b and Supplementary Fig. 1). The optimized patterning protocol (Supplementary Fig. 2) is applied to both polystyrene-*block*-poly(ethylene-*ran*-butylene)-*block*-polystyrene (SEBS) (Extended Data Fig. 2 and Supplementary Figs 3, 4) and polyurethane (Supplementary Fig. 5), with solvent resistance achieved afterwards.

Next, in order to pattern stretchable semiconductors, which are usually incompatible with traditional photolithography processes, we developed two strategies. The first is based on an etching process that involves protection by a copper mask, using a fluorinated polymer[27] as the sacrificial layer, the solvent for which will not dissolve other existing components of the device (Fig. 2c and Extended Data Fig. 3). In separating the film deposition from the patterning process, this method should be universally applicable to various stretchable semiconductors without sacrificing electrical performance (Extended Data Fig. 4). The second strategy uses inkjet printing[28] as an additive patterning method for large-scale fabrication (Fig. 2d and Extended Data Fig. 5), and is applicable to polymer semiconductors that show good solubility.

Using these processes, we fabricated a sheet comprising an intrinsically stretchable array of 108 transistors (Figs 1b and 2e, f), using crosslinked SEBS as the dielectric, a 'conjugated polymer/elastomer phase separation induced elasticity' (CONPHINE) film[16] (patterned by the etching method) as the semiconductor, and carbon nanotubes[22] (spray-coated onto the device through shadow masks—thin metal plates with openings to generate patterns) as the electrodes (Supplementary Fig. 6). Here, to enable higher mobility, we modified the surface energy of the patterned SEBS dielectric layer to obtain the desirable CONPHINE film morphology (Extended Data Fig. 6). As shown in Fig. 2f, the transistors occupied most of the area and achieved a rather high density of 347 per cm^2. In a magnified image of one transistor in the array (Fig. 2g), all of the fully patterned and well aligned components can be clearly seen.

Our intrinsically stretchable transistor array devices show ideal switching behaviour, with no current hysteresis, an on/off current ratio as high as 10^4, minimal gate leakage (Fig. 3a and Supplementary Fig. 7), and good on-shelf and bias stabilities (Supplementary Fig. 8). Furthermore, low-voltage operation at around 10 V (Supplementary Fig. 10) makes the devices suitable for on-skin applications. An overall yield of 94.4% and high performance uniformity (Fig. 3b and Supplementary Fig. 11) were achieved for the 108-transistor array, even though it was fabricated in a non-clean-room environment (with just six transistors out of these 108 failing, owing to leakage of the dielectric layer). The charge-carrier mobilities from all of the working devices show a narrow distribution (Fig. 3c), with an average of $0.821 \pm 0.105\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$, and the highest value reaching $1.11\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. For a less-dense array with a larger channel length (110 μm), the mobility is even higher (average $1.37\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$;

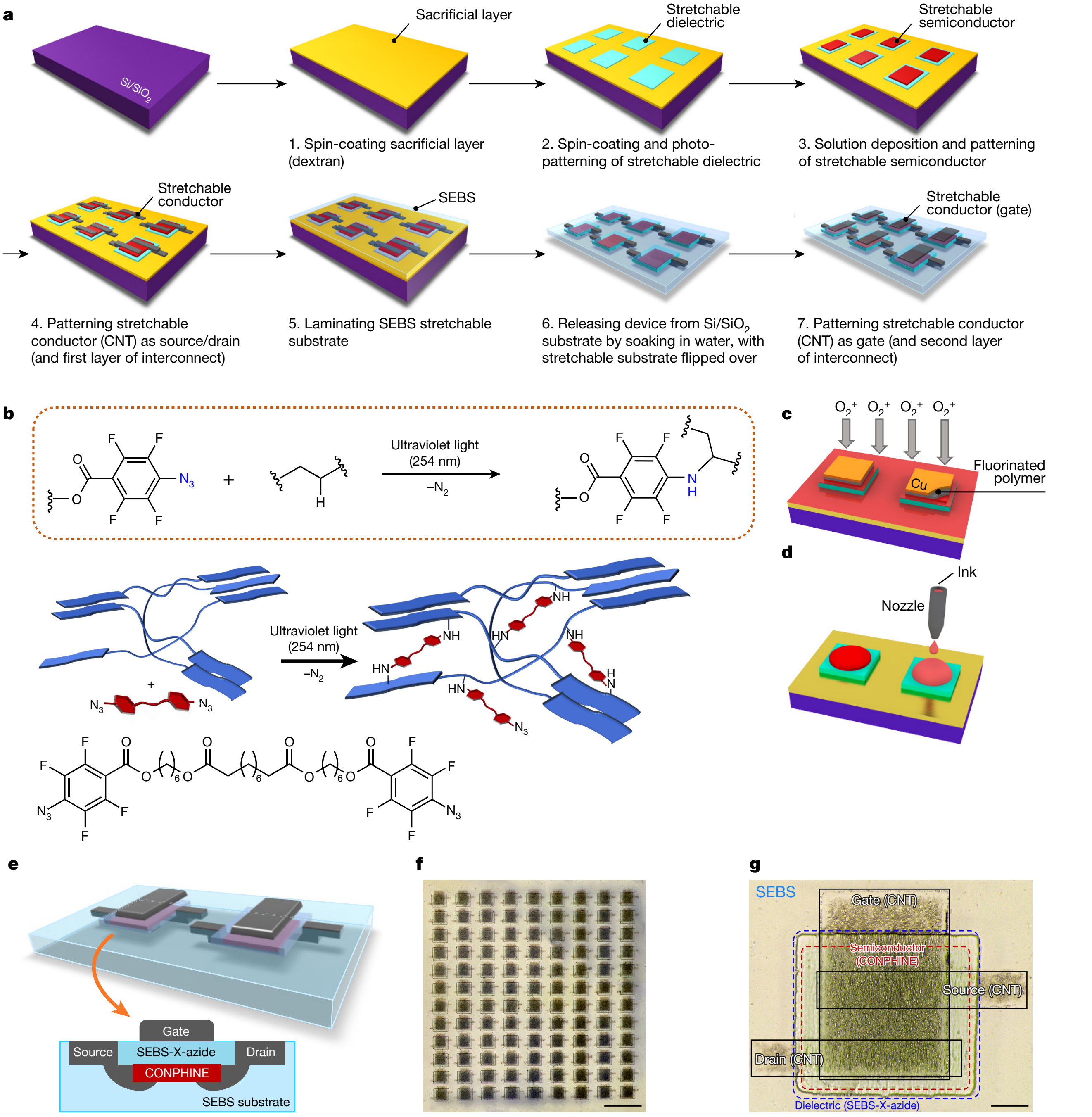

a
Si/SiO$_2$
Sacrificial layer
Stretchable dielectric
Stretchable semiconductor
1. Spin-coating sacrificial layer (dextran)
2. Spin-coating and photo-patterning of stretchable dielectric
3. Solution deposition and patterning of stretchable semiconductor
Stretchable conductor
SEBS
Stretchable conductor (gate)
4. Patterning stretchable conductor (CNT) as source/drain (and first layer of interconnect)
5. Laminating SEBS stretchable substrate
6. Releasing device from Si/SiO$_2$ substrate by soaking in water, with stretchable substrate flipped over
7. Patterning stretchable conductor (CNT) as gate (and second layer of interconnect)
b
F F O O N$_3$ F F
+
H
Ultraviolet light (254 nm)
–N$_2$
N H
NH
HN
N$_3$
c
O$_2^+$ O$_2^+$ O$_2^+$ O$_2^+$
Fluorinated polymer
Cu
d
Ink
Nozzle
e
Gate
Source
SEBS-X-azide
Drain
CONPHINE
SEBS substrate
f
g
SEBS
Gate (CNT)
Semiconductor (CONPHINE)
Source (CNT)
Drain (CNT)
Dielectric (SEBS-X-azide)

Figure 2 | Platform for fabricating intrinsically stretchable transistor arrays. a, Fabrication process flow. We begin with a Si/SiO$_2$ wafer that is coated with a water-soluble sacrificial layer (dextran). Next, a stretchable dielectric is deposited by spin coating and then photo-patterned (see panel **b**). A stretchable semiconductor is deposited by solution deposition and also patterned (see panels **c** and **d**). Stretchable conductors (carbon nanotubes, CNTs) are deposited and patterned as source and drain electrodes, together with the first layer of interconnect. A stretchable substrate (SEBS) is laminated on; soaking in water then releases the device from the rigid substrate. Finally, the gate electrodes (and the second layer of interconnect) are deposited and patterned to complete the transistor structure. **b**, Top, an azide-crosslinking reaction, which is initiated by ultraviolet light and is based on the reaction between azide groups and CH groups. Middle, how the polymer-chain network in an elastomer (blue, with rectangular planes representing rigid segments and tortuous lines representing soft segments) becomes crosslinked by azides (red) into a three-dimensional network. Bottom, chemical structure of an azide crosslinker. Azide crosslinking is a generally applicable strategy for photo-patterning intrinsically stretchable dielectrics. **c**, Etching-based process for patterning stretchable semiconductors. With the protection of patterned copper masks on top of a fluorinated polymer thin film (as the sacrificial layer), the stretchable semiconductor film is patterned by oxygen-plasma treatment. **d**, Inkjet printing as an additive patterning process for stretchable semiconductor film. **e**, Example of an intrinsically stretchable transistor array for performance characterization and demonstration. **f**, Optical microscopic image of a transistor array with 108 transistors. Scale bar, 1 mm. **g**, Magnified image of one transistor in the array. SEBS-X-azide, azide-crosslinked SEBS. Scale bar, 100 µm; channel length, 70 µm; channel width, 270 µm; gate dielectric thickness, 1.25 µm, capacitance, 1.75 nF cm^{-2}; semiconductor thickness, around 130 nm.

maximum 1.78 cm^2 V^{-1} s^{-1}) (Supplementary Fig. 12), showing the possibility of further improving the mobility by just improving the source/drain contacts. In addition, the power consumption of our intrinsically stretchable transistors, in the tens of microwatts, suggests the possibility of self-powered operations[29] for skin electronics.

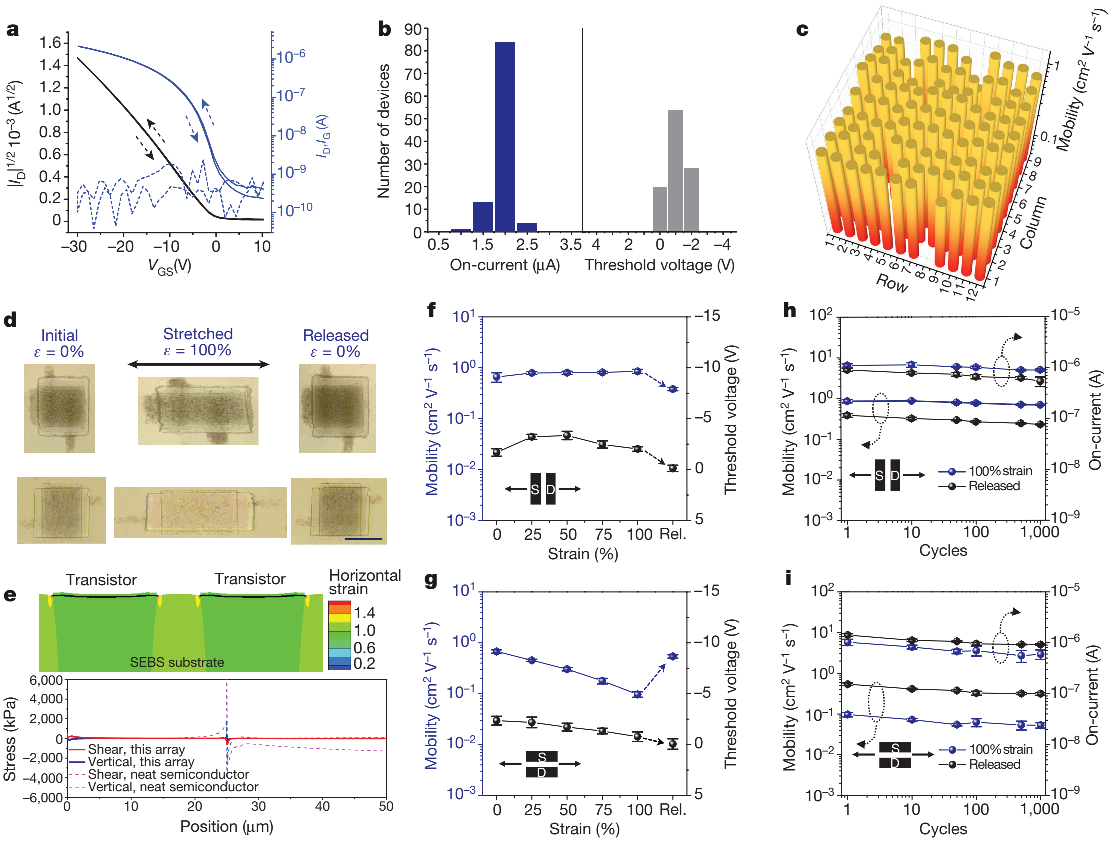


Figure 3 | Electrical performance and stretchability of the intrinsically stretchable transistor array. a, Typical transfer characteristics from the transistor array without strain showing little current hysteresis. I_D, drain current, represented by solid blue and black lines; I_G, gate current, represented by the dashed blue line; V_{GS}, the gate–source voltage. The applied drain–source voltage, V_{DS}, is −30 V. **b**, Histograms showing on-currents and threshold voltages from the 102 working transistors in the 108-transistor array. **c**, Map showing charge-carrier mobility for each transistor location. **d**, Optical microscope images showing the arrayed transistors stretched from 0% to 100% strain (*ε*) and then released, in directions both parallel (top) and perpendicular (bottom) to the channel. Scale bar, 250 μm. **e**, Top, mechanical simulations showing the strain distribution in the array when stretched to 100% strain (with the vertical dimension exaggerated by 25 times). Bottom, the shear stress and vertical stress distribution (under 40% global strain) at the dielectric layer's bottom interface ('this array'), and for a control structure in which the CONPHINE semiconductor layer was replaced with a neat conjugated polymer. 0 represents the dielectric edge and 25 μm represents the semiconductor edge. **f**, **g**, Mobilities and threshold voltages during a stretching cycle parallel (**f**) and perpendicular (**g**) to the channel direction. 'Rel.' refers to the values after releasing from 100% strain. S, source; D, drain. **h**, **i**, Mobilities and on-currents obtained under 100% strain and in a released state during 1,000 stretching cycles, parallel (**h**) and perpendicular (**i**) to the channel direction. The dashed circles and arrows point to the relevant *y*-axis. All data points in **f–i** represent average values obtained from five transistor devices for each set of stretching tests; error bars represent standard deviations.

Our transistor array can be stretched to 100% strain both parallel and perpendicular to the direction of charge transport, with no cracks, delamination or wrinkles observed (Fig. 3d). This is enabled by the 'quasi-homogeneous' mechanical structure design. With all of the major thin-film components being primarily SEBS based (that is, the dielectric, the substrate and 70% of the CONPHINE semiconductor), the interface shear and vertical stresses—the major causes of delamination under stretching—are effectively suppressed (Fig. 3e and Supplementary Fig. 13). The electrical performance of the transistors under channel-parallel stretching is highly stable even at 100% strain (Supplementary Fig. 14a, c), and even shows a slight increase in mobility to 0.99 cm^2 V^{-1} s^{-1} (Fig. 3f), probably owing to the strain-induced alignment of the conjugated polymer fibres in the CONPHINE film[16]. Upon releasing, the change in mobility possibly comes from the change in interface contact resulting from stretching-induced local plastic deformation, and also from the slight viscoelasticity of SEBS. Measured along the direction perpendicular to stretching, a highly stable transfer curve is maintained up to 100% strain (Supplementary Fig. 14b, d), while the slight decrease in mobility (Fig. 3g) agrees with reported characteristics for the CONPHINE film[16]. This intrinsically stretchable transistor array also shows unprecedented mechanical robustness when stretched repeatedly to 100% strain for 1,000 cycles in both directions (Fig. 3h, i and Supplementary Fig. 15). Furthermore, the array is exceptionally stretchable (up to 600% strain), and shows highly stable electrical performance even when subjected to pressure, twisting and biaxial stretching (Supplementary Fig. 16).

The high yield and performance uniformity of this fabrication platform has thus allowed us to develop intrinsically stretchable basic circuit elements (Fig. 1d)—the core of skin electronics. By incorporating scan and data interconnect lines, we achieved an intrinsically stretchable active matrix, with the same device density of 347 transistors per cm^2 (Fig. 4a). As a proof-of-concept demonstration of its application as the multiplexing backplane for skin electronics, we integrated this matrix with a 10 × 10 array of intrinsically stretchable resistive tactile sensors that are based on interdigitated carbon-nanotube electrodes (Fig. 4b and Extended Data Fig. 7), with a resolution of one sensor per 2 mm. The high stretchability allows the array to be attached conformably onto a human palm (with its naturally irregular surface and deformation) as a secondary skin (Fig. 4c and Supplementary Fig. 17), on which the location of a small artificial small ladybug with six conductive legs is accurately detected through a matched map of on-current magnitudes from the sensor pixels (Fig. 4d). This demonstrates the feasibility of using our intrinsically

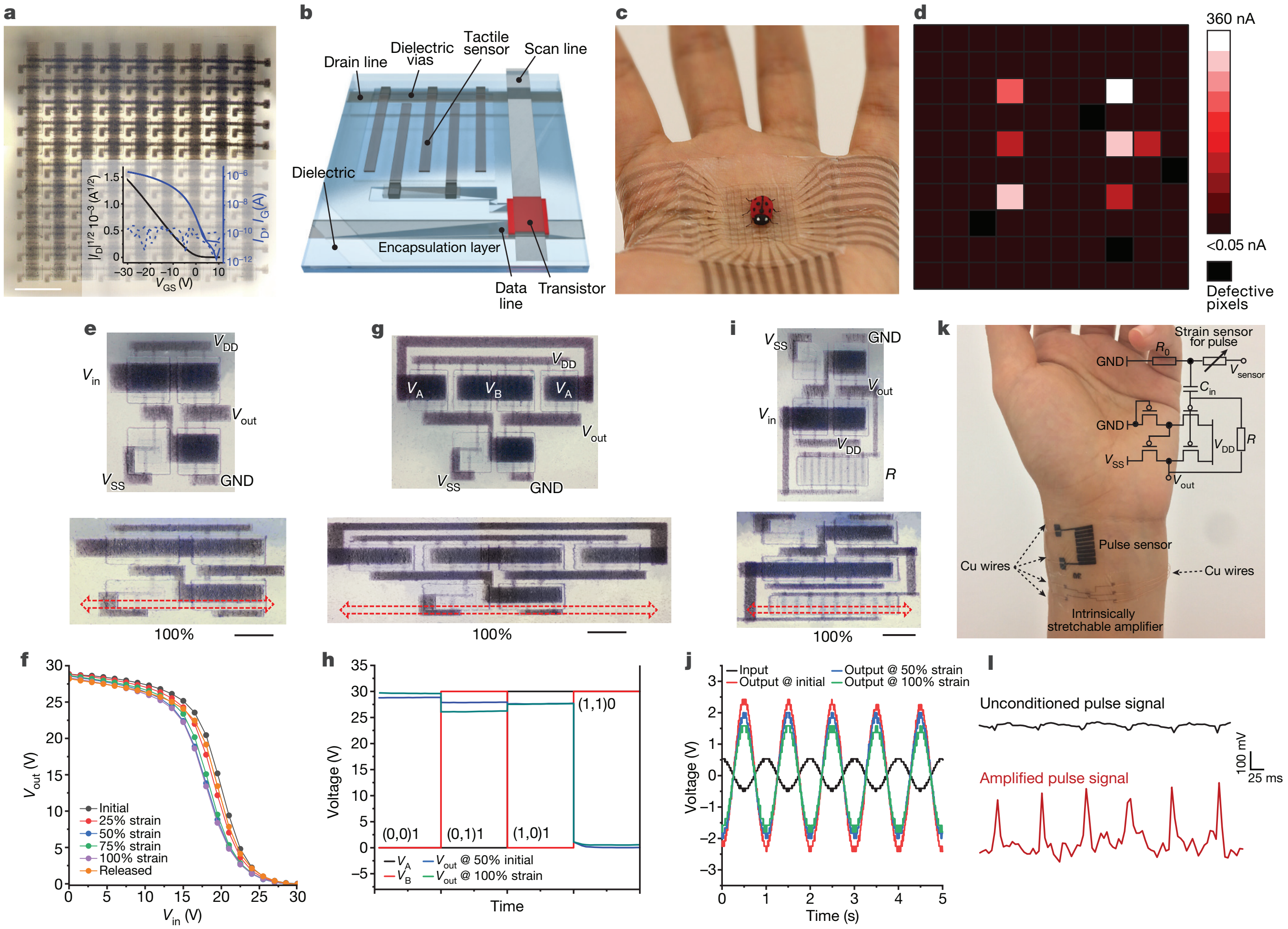


Figure 4 | Intrinsically stretchable circuits for skin electronics. **a**, A stretchable active matrix developed from our intrinsically stretchable transistor array. Scale bar, 1 mm. The inset shows a typical transfer curve recorded from a transistor in the matrix. **b**, Diagram showing a tactile sensor array made from a stretchable active matrix. **c**, The array adheres and conforms to a human palm, enabling accurate sensing of the position of a synthetic ladybug with six conductive legs. Voltages for the scan lines, data lines and drain lines are respectively −10 V, 0 and −10 V during the multiplexing measurement. ('Multiplexing' refers to a method by which one of multiple signals is selected and forwarded to a single channel at a time.) **d**, Current mapping in linear scale, showing exact matching with the position of the ladybug. **e**, Optical microscope images of a fabricated intrinsically stretchable inverter with pseudo-CMOS design, in its initial state (top) and after being stretched to 100% strain (bottom). **f**, Transfer curves from the inverter when stretched gradually from 0% to 100% strain. **g**, Optical microscope images of a fabricated intrinsically stretchable NAND gate in its initial state (top) and after being stretched to 100% strain (bottom). V_A and V_B are the input voltages at the A and B terminals. **h**, Output–input characteristics of the NAND gate at 0% and 100% strain. **i**, Optical microscope images of a fabricated intrinsically stretchable amplifier in its initial state (top) and after being stretched to 100% strain (bottom). **j**, Input sinusoidal signal, along with output signals after amplification when the amplifier is at 0%, 50% and 100% strain. **k**, Use of the intrinsically stretchable amplifier to amplify arterial pulse signals measured by a stretchable strain sensor. The devices are attached on skin side-by-side. Electrical connections are made using Cu wires (as marked), attached on the contacting pads by anisotropic conductive tapes. The inset shows the circuit diagram: R_0 is the divider resistor (680 kΩ); V_{sensor} is the direct-current voltage (40 V) applied to sensor and resistor; and C_{in} is the input capacitor (1 μF) for the amplifier. **l**, Pulse signals obtained, before and after amplification, using the same scale. In all of these circuits, the diode-type transistors have channels of length and width 80 μm; all the other transistors have channels of length and width 80 μm and 1,620 μm. GND, ground. For the electrical characterizations, the direct-current voltages applied to the electrodes indicated are V_{DD} = 30 V and V_{SS} = −30 V. All scale bars represent 600 μm.

stretchable active matrix as the backplane for high-resolution touch sensing in highly conformable electronic skins.

We also realized intrinsically stretchable circuits for signal manipulation and computation, which require higher complexity in terms of transistor interconnection and collective operation. We built basic building-block circuits for digital electronics using this transistor array. For example, an inverter based on a pseudo-CMOS design[30] (Fig. 4e and Extended Data Fig. 8a) gives the expected transfer behaviour, with the same range of input and output voltages, and little circuit-level bias-stress or electrical heating despite continuous operation (Supplementary Fig. 18). When gradually stretched to 100% strain, there is only a small shift in the transfer curve, which is acceptable for logic operations (Fig. 4f). We also built a NAND gate consisting of six transistors (Fig. 4g and Extended Data Fig. 8b), as a 'universal' logic gate for constructing all other logic gates. This shows stable logic operation even under 100% strain (Fig. 4h). For analogue circuits that interface directly with sensors, the basic component is an amplifier circuit. We successfully fabricated such an amplifier using a self-feedback design[31], with a non-gated transistor as the resistor (with resistance of the order of 10^8 Ω; Fig. 4i and Extended Data Fig. 8c); the signal amplification is maintained even at 100% strain (Fig. 4j and Supplementary Fig. 19). Next, to demonstrate skin electronics that contain both sensors and signal-processing units[32], we combined the amplifier with a stretchable pulse sensor[33] (Fig. 4k and Supplementary Fig. 20), achieving on-skin amplification of raw detected physiological signals (Fig. 4l). These basic circuit elements are now ready to be integrated into more complex digital and analogue circuits, to realize more advanced signal-processing functionalities.

Compared with previously reported milestones in developing stretchable transistors and circuits for skin electronics (Extended Data Table 1), our intrinsically stretchable transistor array has, for the first time (to our knowledge), combined advanced electronic functionality with high skin-like stretchability. Moreover, our fabrication process provides a platform on which to readily incorporate future materials advancements into functional electronic circuits and systems with skin-like and even 'beyond-skin' softness and deformability.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 July; accepted 18 December 2017.
Published online 19 February 2018.

1. Gao, W. *et al.* Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature* **529,** 509–514 (2016).
2. Kim, D. H. *et al.* Epidermal electronics. *Science* **333,** 838–843 (2011).
3. Son, D. *et al.* Multifunctional wearable devices for diagnosis and therapy of movement disorders. *Nat. Nanotechnol.* **9,** 397–404 (2014).
4. Tee, B. C. K. *et al.* A skin-inspired organic digital mechanoreceptor. *Science* **350,** 313–316 (2015).
5. Kang, S. K. *et al.* Bioresorbable silicon electronic sensors for the brain. *Nature* **530,** 71–76 (2016).
6. Minev, I. R. *et al.* Electronic dura mater for long-term multimodal neural interfaces. *Science* **347,** 159–163 (2015).
7. Khodagholy, D. *et al. In vivo* recordings of brain activity using organic transistors. *Nat. Commun.* **4,** 1575 (2013).
8. Li, S., Zhao, H. C. & Shepherd, R. F. Flexible and stretchable sensors for fluidic elastomer actuated soft robots. *MRS Bull.* **42,** 138–142 (2017).
9. Rus, D. & Tolley, M. T. Design, fabrication and control of soft robots. *Nature* **521,** 467–475 (2015).
10. Kaltenbrunner, M. *et al.* An ultra-lightweight design for imperceptible plastic electronics. *Nature* **499,** 458–463 (2013).
11. Matsuhisa, N. *et al.* Printable elastic conductors with a high conductivity for electronic textile applications. *Nat. Commun.* **6,** 7461 (2015).
12. Xu, S. *et al.* Soft microfluidic assemblies of sensors, circuits, and radios for the skin. *Science* **344,** 70–74 (2014).
13. Chortos, A. *et al.* Mechanically durable and highly stretchable transistors employing carbon nanotube semiconductor and electrodes. *Adv. Mater.* **28,** 4441–4448 (2016).
14. Chortos, A. *et al.* Highly stretchable transistors using a microcracked organic semiconductor. *Adv. Mater.* **26,** 4253–4259 (2014).
15. Liang, J. J. *et al.* Intrinsically stretchable and transparent thin-film transistors based on printable silver nanowires, carbon nanotubes and an elastomeric dielectric. *Nat. Commun.* **6,** 7647 (2015).
16. Xu, J. *et al.* Highly stretchable polymer semiconductor films through the nanoconfinement effect. *Science* **355,** 59–64 (2017).
17. Oh, J. Y. *et al.* Intrinsically stretchable and healable semiconducting polymer for organic transistors. *Nature* **539,** 411–415 (2016).
18. Scott, J. I. *et al.* Significantly increasing the ductility of high performance polymer semiconductors through polymer blending. *ACS Appl. Mater. Interfaces* **8,** 14037–14045 (2016).
19. Wang, Y. *et al.* A highly stretchable, transparent, and conductive polymer. *Sci. Adv.* **3,** e1602076 (2017).
20. Larson, C. *et al.* Highly stretchable electroluminescent skin for optical signaling and tactile sensing. *Science* **351,** 1071–1074 (2016).
21. Liang, J. J., Li, L., Niu, X. F., Yu, Z. B. & Pei, Q. B. Elastomeric polymer light-emitting devices and displays. *Nat. Photon.* **7,** 817–824 (2013).
22. Lipomi, D. J. *et al.* Skin-like pressure and strain sensors based on transparent elastic films of carbon nanotubes. *Nat. Nanotechnol.* **6,** 788–792 (2011).
23. You, I. *et al.* Stretchable E-skin apexcardiogram sensor. *Adv. Mater.* **28,** 6359–6364 (2016).
24. White, M. S. *et al.* Ultrathin, highly flexible and stretchable PLEDs. *Nat. Photon.* **7,** 811–816 (2013).
25. Png, R. Q. *et al.* High-performance polymer semiconducting heterostructure devices by nitrene-mediated photocrosslinking of alkyl side chains. *Nat. Mater.* **9,** 152–158 (2010).
26. Liu, L. H. & Yan, M. D. Perfluorophenyl azides: new applications in surface functionalization and nanomaterial synthesis. *Acc. Chem. Res.* **43,** 1434–1443 (2010).
27. Choi, W., Kim, M. H., Na, Y. J. & Lee, S. D. Complementary transfer-assisted patterning of high-resolution heterogeneous elements on plastic substrates for flexible electronics. *Org. Electron.* **11,** 2026–2031 (2010).
28. Singh, M., Haverinen, H. M., Dhagat, P. & Jabbour, G. E. Inkjet printing-process and its applications. *Adv. Mater.* **22,** 673–685 (2010).
29. Hwang, B. U. *et al.* Transparent stretchable self-powered patchable sensor platform with ultrasensitive recognition of human activities. *ACS Nano* **9,** 8801–8810 (2015).
30. Huang, T. C. *et al.* Pseudo-CMOS: a design style for low-cost and robust flexible electronics. *IEEE Trans. Electron Dev.* **58,** 141–150 (2011).
31. Sekitani, T. *et al.* Ultraflexible organic amplifier with biocompatible gel electrodes. *Nat. Commun.* **7,** 11425 (2016).
32. Koo, J. H. *et al.* Wearable electrocardiogram monitor using carbon nanotube electronics and color-tunable organic light-emitting diodes. *ACS Nano* **11,** 10032–10041 (2017).
33. Wu, Y. *et al.* High resolution flexible strain sensors for biological signal measurements. In *19th International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers)* 1144–1147 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work is supported by Samsung Electronics. We thank L. Beker, N. Matsuhisa and S. Chen for helping with experiments, and L. Jin for discussions about mechanical simulation. J.L. acknowledges support by the National Science Foundation Graduate Research Fellowship Program under grant DGE-114747. S.-K.K. thanks NETEP and MOTIE of the Republic of Korea (grant 20173010013000).

Author Contributions S.W., J.X. and Z.B. designed the project and experiments. S.W., J.X. and W.W. fabricated the intrinsically stretchable transistor array and circuits, and carried out electrical characterizations. G.-J.N.W. synthesized the azide compound. R.R. carried out mechanical simulations. S.W. and F.M.-L. undertook the inkjet printing of a semiconducting polymer. V.R.F. carried out X-ray photoelectron spectroscopy characterizations. J.W.C., A.M.F. and A.E. helped to prepare schematics for the three-dimensional transistor array and carried out device photography. J.X. and J.L. did the mechanical characterizations. S.-K.K. and A.G. provided conjugated polymers. T.L. helped with development of semiconductor patterning. S.N., Y.K., Y.Y. and B.M. helped with circuit design and measurements. S.W., Z.B., J.X. and J.B.-H.T. wrote the manuscript. All authors reviewed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Z.B. (zbao@stanford.edu).

METHODS

Materials. All processing solvents, such as chlorobenzene, toluene, dodecane, hexane, 1,2,3,4-tetrahydronaphthalene (tetralin), 2-propanol, and ethoxynonafluorobutane, were purchased from commercial sources and used as received. The polymer semiconductor poly-[2,5-bis(7-decylnonadecyl)pyrrolo[3,4-c]pyrrole-1,4-(2H,5H)-dione-(E)-(1,2-bis(5-(thiophen-2-yl)selenophen-2-yl)ethene) (29-DPP-SVS) was synthesized as reported[34], as was poly(thiophen-2-yl)ethyl-2,5-bis(4-decyl tetradecyl)-2,5-dihydro-pyrrolo[3,4-c]pyrrole-1,4-dione-r-10%-2,6-pyridine-dicarboxamide (DPP2TTVT-PDCA)[17]. The number-averaged molecular weight and polydispersity of these polymers are shown in Supplementary Table 1. Number-average molecular weight (M_n), weight-average molecular weight (M_w), and polydispersity index (PDI) were evaluated by high-temperature size exclusion chromatography (SEC) using 1,2,4-trichlorobenzene and performed at 180 °C with an EcoSEC high-temperature gel permeation chromatography (GPC) device (Tosoh). This is equipped with a single TSKgel GPC column (GMH$_{HR}$-H; 300 mm × 7.8 mm), which was calibrated by monodisperse polystyrene standards. The azide crosslinker bis(6-((4-azido-2,3,5,6-tetrafluorobenzoyl)oxy)hexyl) decanedioate was synthesized as described below. The SEBS compounds H1221 and H1052, with poly(ethylene-co-butylene) volume fractions of respectively 88% and 80%, were provided by the Asahi Kasei company. SEBS rubbers have been reported to have good long-term stability and biocompatibility[35,36]. We use SEBS H1052 as a stretchable dielectric layer, and SEBS H1221 in the CONPHINE semiconductor and also as the stretchable substrate in the intrinsically stretchable transistor array. PU SG80A was supplied by Lubrizol. Dextran was purchased from Sigma-Aldrich and used as received. Carbon nanotubes for electrodes were purchased from Carbon Solutions (P3-SWNTs). Methyl pentafluorobenzoate and sebacoyl chloride were purchased from Oakwood Chemical and ACROS Organics, respectively, and used as received. Other commercial reactants were purchased from Sigma-Aldrich and used without further purification.

Fabrication of the intrinsically stretchable transistor array. As the substrate for the fabrication process, a Si/SiO$_2$ wafer was first cleaned with oxygen plasma (150 W, 200 mTorr) for 2 min, and then sonicated in acetone, 2-propanol and deionized water for 5 min each. After the wafer was blown dry with nitrogen gas, a dextran solution (10 wt% in water) was spin-coated on top at 1,500 r.p.m. for 20 s. The wafer was then baked on a hot plate at 80 °C for 1 min and 180 °C for 30 min to fully remove the trapped water in the film. Then, a solution containing SEBS (H1052) with azide crosslinker (60 mg ml^{-1} for SEBS and 2.4 mg ml^{-1} for azide in toluene) was spin-coated on top at 1,000 r.p.m., to form a stretchable dielectric film with a thickness of about 1.25 μm. Subsequently, the SEBS–azide film was photo-patterned by exposure under deep ultraviolet light (wavelength 254 nm, using a Spectrum 100 Precision UV Spot Curing System from American Ultraviolet) for 4 min with a dose of about 540 mJ cm^{-2}, to initiate the photo-crosslinking reaction in selective areas defined by a mask. Soft baking was carried out at 120 °C for 15 min in air on a hot plate to further increase the degree of crosslinking in photo-exposed areas. After this, dodecane was used to dissolve the unexposed areas of SEBS, with the exposed areas preserved. In order to fully crosslink these preserved areas, the device was further baked at 200 °C for 1 h in a glovebox.

The surface of the patterned SEBS ('SEBS-X-azide') dielectric was modified by octadecyltrimethoxysilane (OTS) molecules to increase the hydrophobicity, through consecutive processes of O$_2$ plasma treatment, spin-coating of OTS solution (3 mM in hexane) at 3,000 r.p.m., and finally vapour annealing in a desiccator with a small vial containing a few millimetres of ammonium hydroxide solution (28–30% in water) for 10 h at room temperature. Then, the CONPHINE semiconductor film was deposited on top by spin-coating of 29-DPP-SVS-(1)/SEBS-H1221 solution (10 mg ml^{-1} in chlorobenzene, with a weight ratio of 3:7) at 1,000 r.p.m., followed by annealing at 150 °C on a hotplate in a glovebox. A fluorinated polymer solution (3M Novec 1902 Electronic Grade Coating diluted by ethoxynonafluorobutane in a 1/2 volume ratio) was spin-coated at 1,000 r.p.m. to give the fluorinated sacrificial layer. Copper patterns (200 nm) as etching masks were deposited on top by thermal evaporation, through a shadow mask aligned with the stretchable dielectric pattern. Etching of the CONPHINE semiconductor film in the areas uncovered by the patterned copper films was performed in O$_2$ plasma (150 W, 200 mTorr). Then, the copper films were lifted off by soaking the device in ethoxynonafluorobutane to dissolve the fluorinated polymer sacrificial layer, only leaving the CONPHINE semiconductor patterns on top of the SEBS dielectric patterns. Top source/drain electrodes were patterned by spray-coating carbon nanotube (CNT) solution through a shadow mask (Invar). The CNT solution is prepared by dispersing 20 mg P3-SWNT in 70 ml 2-propanol with two drops of water, through consecutive 3-h bath sonication, 10-min tip sonication and then centrifugation. In order to improve the resolution of the spray-coated CNT electrodes, we placed a piece of magnet underneath the sample to hold the shadow mask tightly to the sample, and optimized the spray-coating condition to the sample temperature of about 75 °C, the flow rate of 2.8 ml min^{-1}, and the gun-to-sample distance of about 12 cm.

In the next fabrication step, a SEBS (H1221) stretchable substrate was conformably laminated onto the fabricated array devices on the Si substrate. Then, the array was transferred onto the stretchable substrate by immersing the entire device in water to dissolve the sacrificial dextran layer. Finally, the gate electrodes were patterned on top of the SEBS dielectric by spray-coating the above-described CNT solution through a shadow mask. The fabricated intrinsically stretchable transistor array was baked at 80 °C under vacuum for 4 h to fully remove the moisture, before electrical characterization. Alignment of the shadow masks was performed under optical microscopy.

Photo-patterning of polyurethane using azide crosslinking. A polyurethane (SG80A) solution with azide crosslinker (40 mg ml^{-1} polyurethane; 1.6 mg ml^{-1} azide in tetrahydrofuran) was spin-coated at 1,000 r.p.m. onto a Si substrate, to form a stretchable dielectric film with a thickness of about 1.2 μm. The polyurethane–azide film was then photo-patterned by exposure under deep ultraviolet light (wavelength 254 nm) for 9 min with a dose of around 1,215 mJ cm^{-2}, to initiate the photo-crosslinking reaction in selective areas defined by the mask. The film was then soft baked at 120 °C for 15 min in air, and developed in *n*-methyl-2-pyrrolidon (NMP), before further baking at 200 °C for 1 h in a glovebox.

Surface modification of patterned SEBS dielectric. The morphology of a CONPHINE film obtained by spin-coating of the solution depends strongly on the surface energy of the substrate. When CONPHINE film is spin-coated on crosslinked SEBS, the SEBS phase in the CONPHINE film is strongly attracted by the dielectric surface, expelling almost all of the conjugated polymers to the top surface of the film (as shown by X-ray photoelectron spectroscopy (XPS) in Extended Data Fig. 6e). The strong mutual interaction in the conjugated polymer phase leads to the formation of a dense network with many large aggregates and junctions (Extended Data Fig. 6b). This morphology might create excess boundaries for charge transport at the aggregates and junctions, and leads to a relatively low charge mobility of 0.31 cm^2 V^{-1} s^{-1} (Extended Data Fig. 6g). In order to turn this morphology into the nanoscale-confined fibre network that favours efficient charge transport, we modify the crosslinked SEBS surface with OTS molecules to increase the hydrophobicity and thereby reduce the attraction to the SEBS phase in the CONPHINE film. Through such surface modification, the water contact angle of crosslinked SEBS increases from 94° to 102° (Extended Data Fig. 6a, b), indicating decreased surface energy. The CONPHINE film spin-coated on this OTS-modified crosslinked SEBS now produces a morphology with a gradually reduced distribution of the conjugated polymer phase throughout the CONPHINE film from the top surface to the bottom surface. The conjugated polymer phase at the top surface forms the ideal nanoconfined morphology without observable large aggregates (Extended Data Fig. 6d). The charge-carrier mobility thereby increases to 0.76 cm^2 V^{-1} s^{-1} (Extended Data Fig. 6h).

Inkjet printing of polymer semiconductor. Inkjet printing was done using a Dimatix Materials printer (DMP-2850). The polymer semiconductor 29-DPP-SVS-(*2*) (M_n = 30.1 kDa) was dissolved in tetralin, producing a concentration of 3 mg ml^{-1}. After filtering, the solution was filled into the printing cartridge, and then desiccated for 20 min. During printing onto SEBS dielectric patterns, the drop-to-drop distance was set to 30 μm, with two adjacent nozzles used at the same time. The temperature of the sample platen was set to 40 °C. The polymer semiconductor patterns aligned on SEBS patterns were obtained using five layers of consecutive printing in order to get enough polymer materials deposited for good performance. Annealing at 150 °C for 30 min in a glovebox, and for another 30 min in a vacuum, led to better crystallization and full removal of the solvent. To evaluate performance, gold was deposited through the aligned shadow mask, as the top contact source/drain electrodes for constructing thin-film transistor structures.

The stretchability of the 29-DPP-SVS-(2) patterns obtained by inkjet printing was compared with that of the 29-DPP-SVS-(2) spin-coated films patterned by copper-protected etching. For both processes, the 29-DPP-SVS-(2) semiconductor was patterned on top of azide-X-SEBS film on dextran-coated Si/SiO$_2$ substrate. Then, the 29-DPP-SVS-(2) patterns together with the SEBS dielectrics were transferred to polydimethylsiloxane (PDMS) stamps by dissolving the dextran layer in water. After stretching to 100% strain, the 29-DPP-SVS-(2) patterns on the azide-X-SEBS dielectric film were transferred to a Si/SiO$_2$ substrate. Finally, gold electrodes were evaporated on top to allow mobility measurements. Initial mobilities were also obtained after the same transferring process to Si/SiO$_2$ substrates in order to maintain comparative experimental conditions.

Fabrication of an active-matrix tactile sensor array. The active-matrix part was fabricated in a similar way to the intrinsically stretchable transistor array, with the same fabrication conditions for all corresponding material components. A Si/SiO$_2$ wafer was used as the substrate, with dextran solution (10 wt% in water) spin-coated as the sacrificial layer at 1,500 r.p.m for 20 s. It was dried by baking at 80 °C for 1 min, and then baked at 180 °C for 30 min to fully remove any trapped water. Next, a thin film of SEBS was photo-patterned through azide-crosslinking chemistry, forming the dielectric layer with via-openings that allow the electrical connection between transistor and tactile sensor. A CONPHINE film as the semiconductor layer was spin-coated on top and patterned by the etching-based strategy described earlier. Carbon nanotubes were then spray-coated through a shadow mask to form stretchable data lines, source and drain electrodes for the transistors, and drain lines (Fig. 4b). A thin layer of SEBS substrate was laminated on top for transferring the device; soaking in water then dissolved the dextran sacrificial layer. On the reverse of the device, scan lines (as gate electrodes for the transistor) and interdigitated electrodes for the tactile sensors were fabricated by spray-coating of carbon nanotubes through a shadow mask.

An encapsulation layer that exposes just the tactile sensor elements was fabricated by photo-patterning a SEBS thin film (spin-coated from a solution of 80 mg ml^{-1} SEBS and 3.2 mg ml^{-1} azide in toluene) through azide-crosslinking chemistry, on top of a dextran-coated Si/SiO$_2$ substrate. The active-matrix tactile sensor array described above, supported by a glass slide, was laminated onto this encapsulation layer, with alignment achieved under an optical microscope. Finally, the entire device was released from the Si/SiO$_2$ substrate by soaking in water to dissolve the dextran. Baking at 80 °C under a vacuum for 4 h fully removed any moisture before electrical characterization.

Fabrication of a resistive strain sensor for pulse monitoring. A 50-μm-thick commercial polyimide tape was attached to a silicon wafer that served as a large, flat, and heat-conductive substrate. Laser-induced porous graphite structures were patterned by direct writing with a CO$_2$ laser on the polyimide film, as described[33]. The SEBS/toluene solution was poured onto the patterned sample, and the solvent allowed to evaporate gradually. After the SEBS film (thickness 0.1 mm) had solidified fully, the graphite strain sensor could be peeled off by hand from the polyimide tape and transferred onto the SEBS film. Finally, to protect the porous graphite pattern from peeling off when making contact with an object, the whole structure was sealed with another thin PDMS layer of about 0.05 mm in thickness.

Material and device characterizations. Mechanical strain–stress and cyclic stress–strain experiments of crosslinked and uncrosslinked SEBS films were performed using an Instron 5565 instrument. Creep recovery was tested through dynamic mechanical analysis (using a TA Instrument Q800 with tension clamps). Crosslinked SEBS films were spin-coated from 150 mg ml^{-1} toluene solution at a speed of 300 r.p.m. for 3 min. The spin-coated films were then laminated several times to form a thick film (about 0.05 mm thick) for clamping. The thick film was further crosslinked at 200 °C for 1 h in a glovebox. An SEBS sample film was prepared in the same way. These films were then used for strain–stress and creep-recovery experiments.

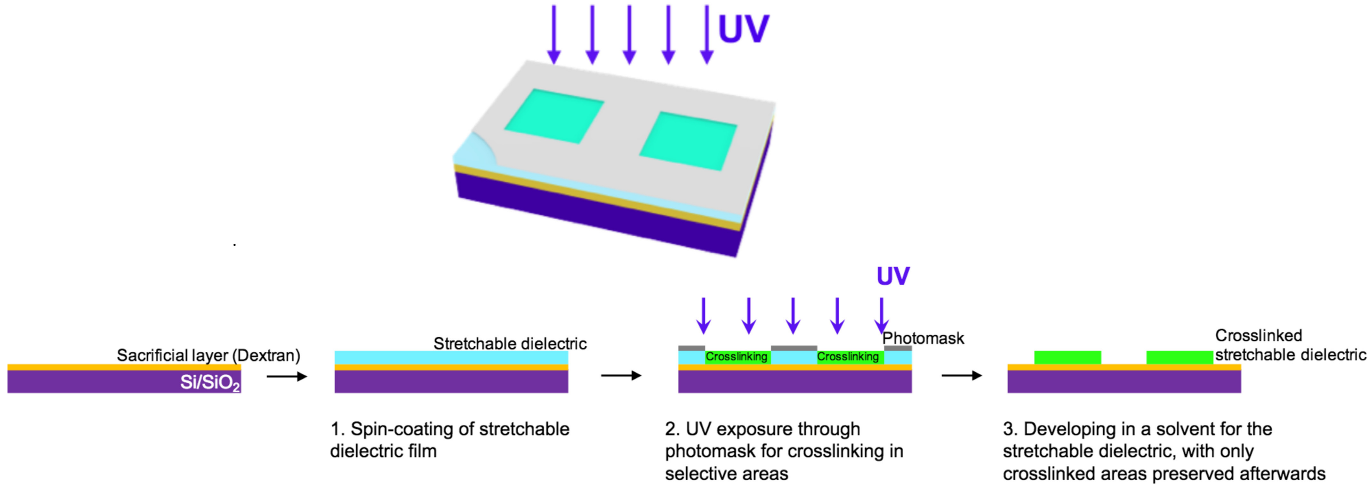
X-ray photoelectron spectroscopy (XPS) characterization was performed using a PHI VersaProbe III scanning XPS microprobe. For depth profiling, sputtering was carried out using an argon 2500+ gas cluster ion beam gun at 5 kV and 15 mA, with 1 min of sputtering per cycle. An SEBS film of known thickness was used to calibrate the sputtering rate, which was 3 nm min^{-1}. The dielectric constants of SEBS before and after azide crosslinking were obtained through capacitance measurement with an Agilent precision LCR meter E4980A. Nuclear magnetic resonance (NMR) spectra were recorded on a Varian Mercury console spectrometer (^{1}H 400 MHz, ^{19}F 400 MHz, ^{13}C 100 MHz). Chemical shifts are given in parts per million (p.p.m.) with respect to tetramethysilane as an internal standard, and coupling constants (J) are given in hertz (Hz).

All of the electrical characteristics of the transistors were measured using a probe station in an ambient environment connected to a Keithley 4200 characterization system. Characterizations of the circuits were also performed using the probe station. The inverter and the NAND circuits were characterized using the Keithley 4200 coupled to an extra direct-current power supply. The amplifier circuit was characterized using a functional generator (Velleman PCSGU250) to generate the sinusoidal input signal and an oscilloscope (Velleman PCSGU250) with a 10/1 probe to measure the output signal, with a AC-coupling capacitor (2 μ F) added at both input and output ends. The active-matrix tactile sensor array on a human palm was electrically connected to external copper wires using anisotropic conductive tapes. Static tactile mapping of the artificial ladybug was performed with the Keithley 4200 using a manual switch.

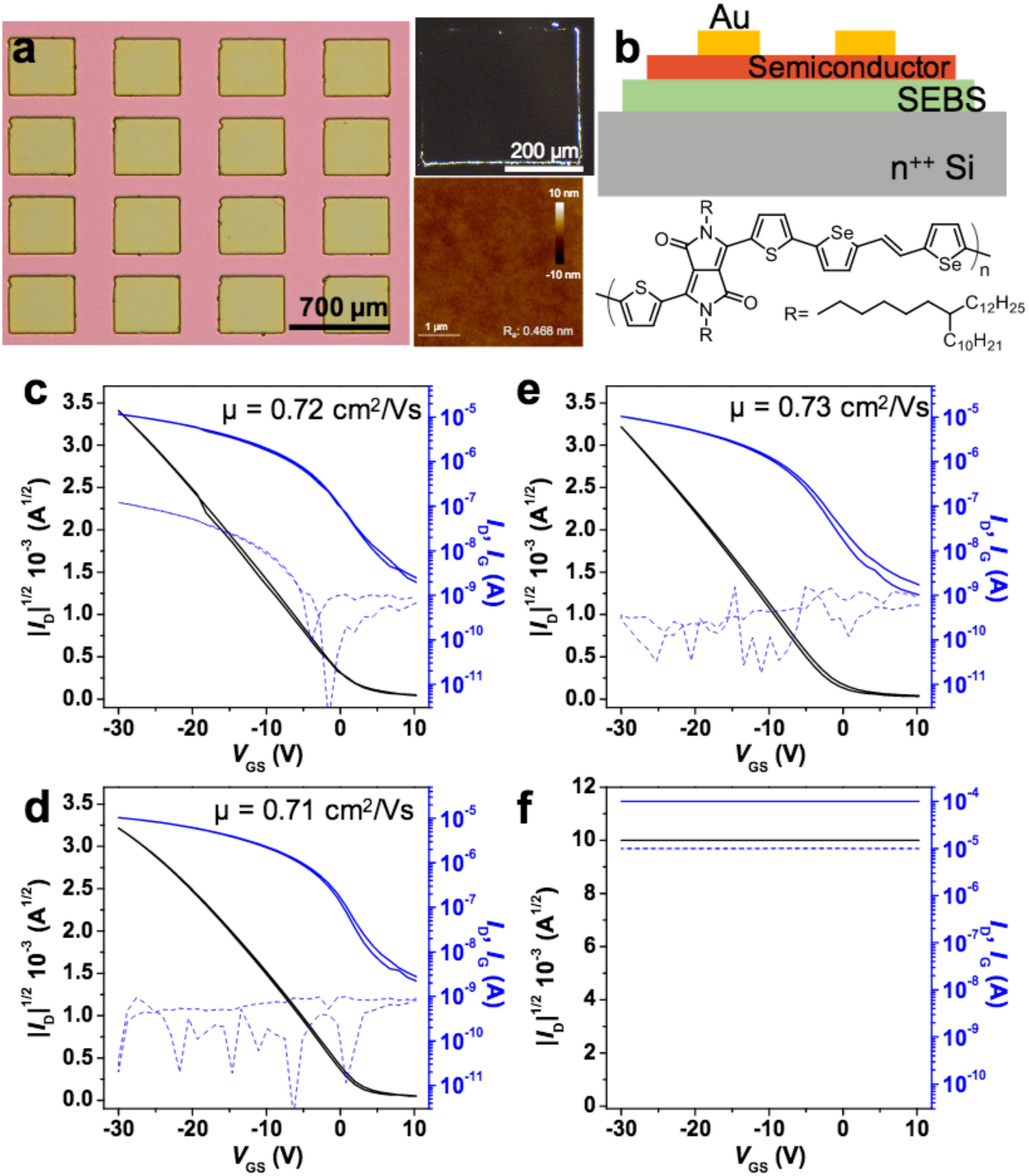
Testing the devices on human hands. The tests of devices on human hands described herein do not need Institutional Review Board (IRB) approval, because our experiments do not affect living people physically or physiologically, and we have not sought or received identifiable private information. The hands shown in Figs 1, 4 and Supplementary Figs 17, 20 are those of J. Xu and S. Wang, who have given their consent to publish these images.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

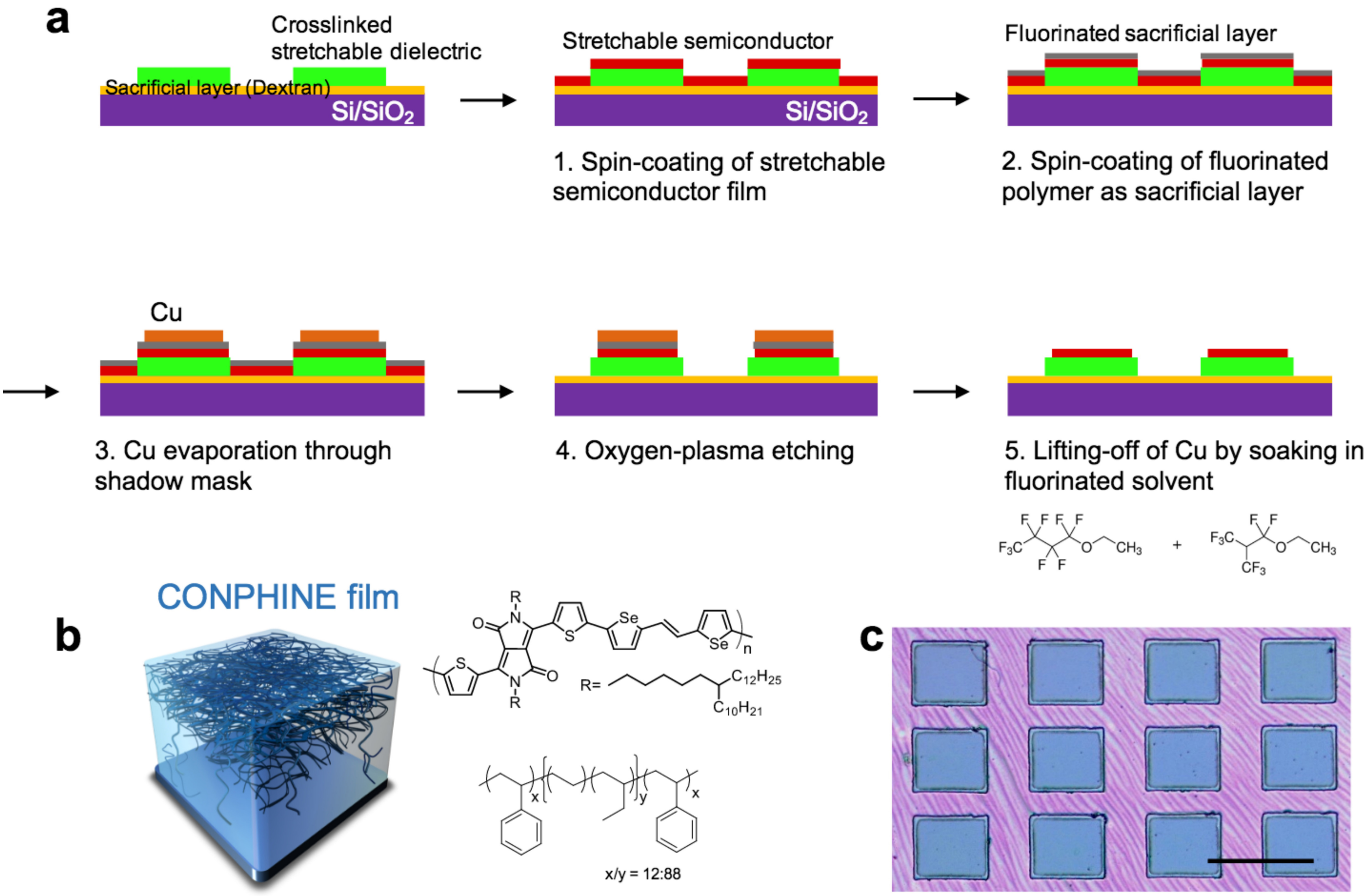
34. Kang, I., Yun, H. J., Chung, D. S., Kwon, S. K. & Kim, Y. H. Record high hole mobility in polymer semiconductors via side-chain engineering. *J. Am. Chem. Soc.* **135,** 14896–14899 (2013).
35. Guillemette, M. D., Roy, E., Auger, F. A. & Veres, T. Rapid isothermal substrate microfabrication of a biocompatible thermoplastic elastomer for cellular contact guidance. *Acta Biomater.* **7,** 2492–2498 (2011).
36. Borysiak, M. D. *et al.* Simple replica micromolding of biocompatible styrenic elastomers. *Lab Chip* **13,** 2773–2784 (2013).
37. Kim, D.-H. *et al.* Stretchable and foldable silicon integrated circuits. *Science* **320,** 507–511 (2008).
38. Sekitani, T. *et al.* A rubberlike stretchable active matrix using elastic conductors. *Science* **321,** 1468–1472 (2008).
39. Cai, L., Zhang, S., Miao, J., Yu, Z. & Wang, C. Fully printed stretchable thin-film transistors and integrated logic circuits. *ACS Nano* **10,** 11459–11468 (2016).


UV
Sacrificial layer (Dextran)
Si/SiO$_2$
Stretchable dielectric
1. Spin-coating of stretchable dielectric film
UV
Photomask
Crosslinking
Crosslinking
2. UV exposure through photomask for crosslinking in selective areas
Crosslinked stretchable dielectric
3. Developing in a solvent for the stretchable dielectric, with only crosslinked areas preserved afterwards

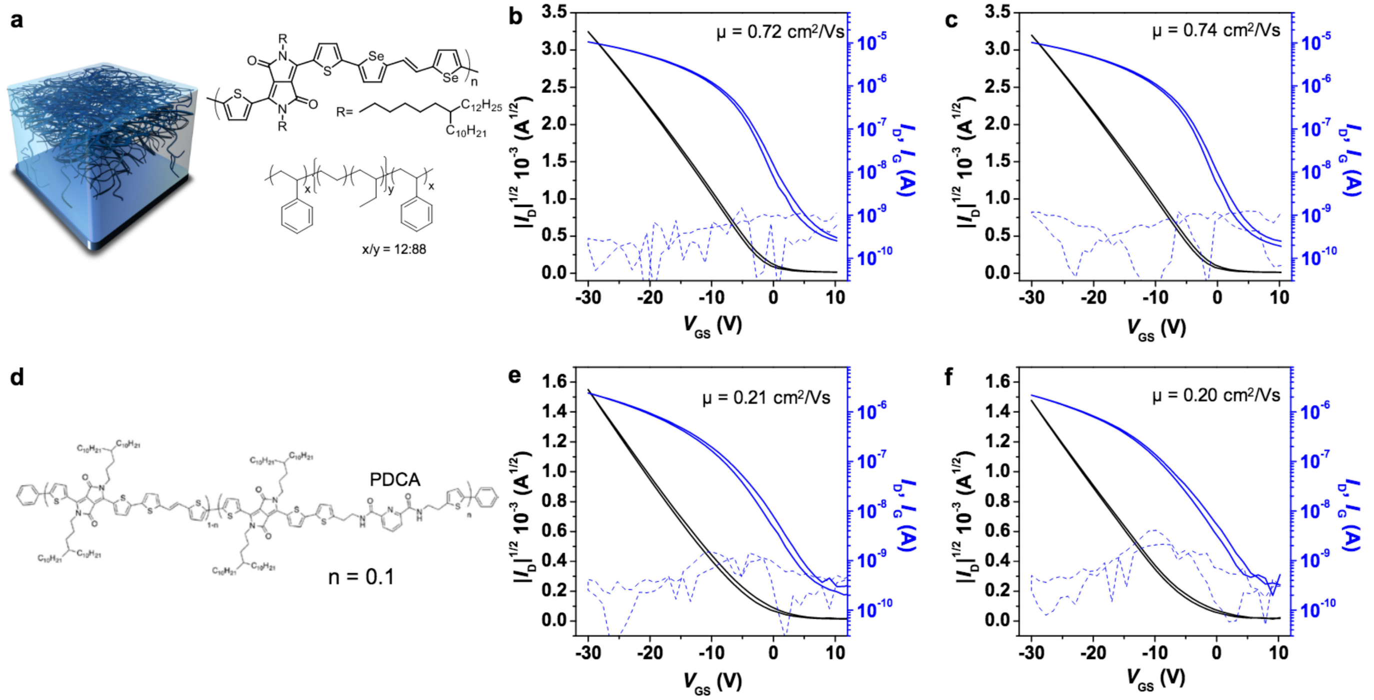
Extended Data Figure 1 | A direct photo-patterning process for fabricating stretchable dielectrics. Top, diagram of the stretchable dielectric undergoing ultraviolet-triggered azide crosslinking through a mask (grey). Bottom, detailed steps of the process.



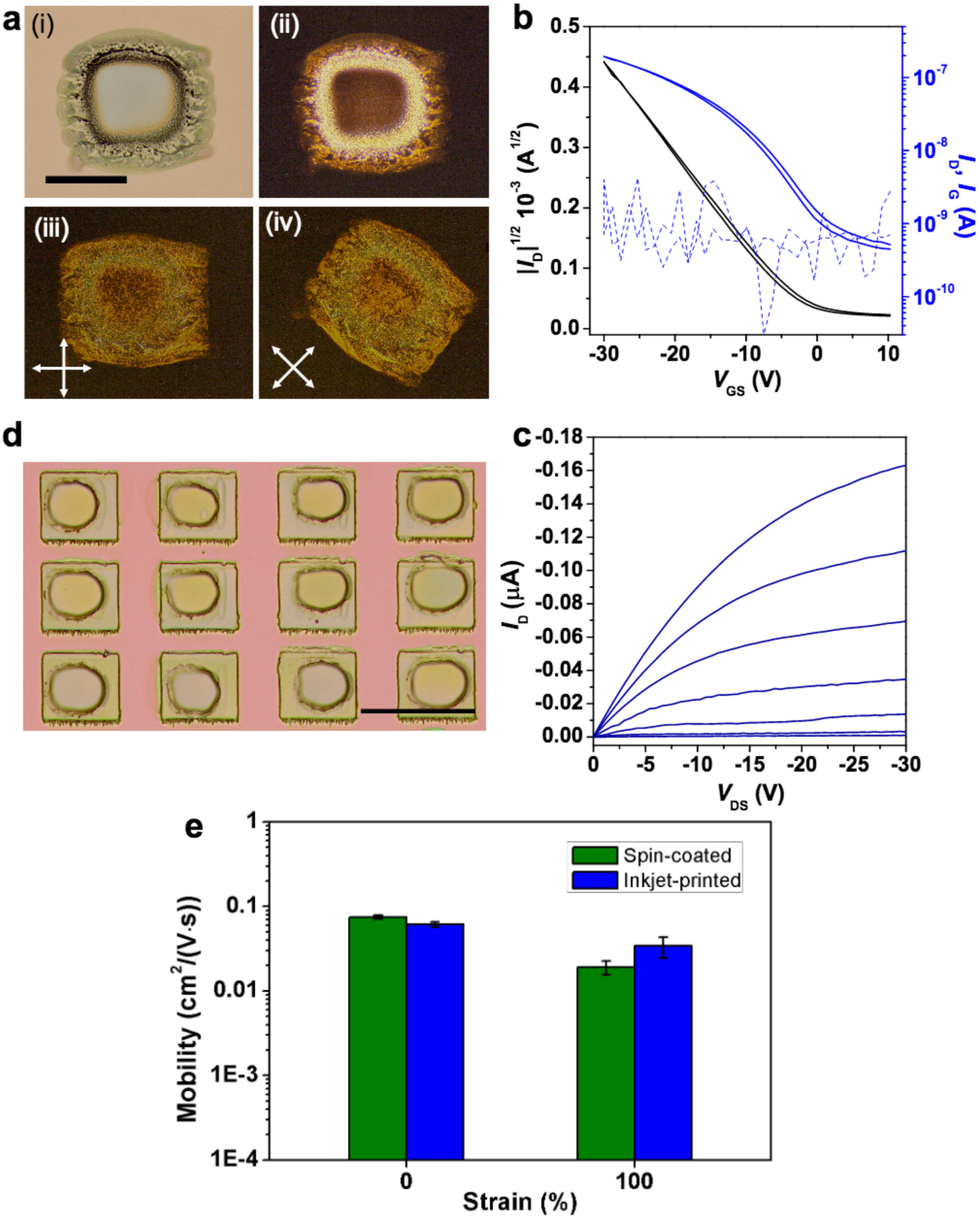
Extended Data Figure 2 | A photo-patterned SEBS film constitutes a solvent-resistant, stretchable dielectric layer for organic thin-film transistors. a, Optical microscope images (left, bright field; top right, dark field) of a photo-patterned SEBS film. Bottom right, an image of the surface taken using an atomic-force microscope (AFM), showing the very small roughness value ($R_a = 0.468$ nm) produced after patterning. **b**, Top, diagram showing an azide-crosslinked SEBS film with semiconductor on top, constituting an organic field-effect transistor. Bottom, chemical structure of the semiconductor layer in the diagram above. We made these transistors to test the solvent resistance of the azide-patterned SEBS film, to ensure that it could allow direct spin-coating of semiconductor solution (with chlorobenzene as the solvent) on top. Channel length, 50 μm; channel width, 1,000 μm; gate dielectric capacitance, 1.75 nF cm^{-2}. **c**, Transfer curve of a transistor that was obtained by transferring semiconductor film onto azide-patterned SEBS. In this case, the semiconductor film was obtained by spin-coating its solution onto an OTS-treated SiO_2 surface.**d**, Transfer curve for a transistor that was obtained by directly spin-coating the semiconductor onto azide-patterned SEBS. The similar mobility (μ) in **c** and **d** indicates that the semiconductor can be spin-coated directly onto the solvent-resistant, azide-patterned SEBS. **e**, Transfer curve for a transistor with the semiconductor film transferred onto unpatterned (uncrosslinked) SEBS. This semiconductor film was obtained by spin-coating its solution onto an OTS-treated SiO_2 surface. The similar mobility in **c** and **e** indicates that azide crosslinking does not change the ability of SEBS to function as a dielectric. **f**, Transfer curve for a transistor with the semiconductor film directly spin-coated onto unpatterned (uncrosslinked) SEBS, during which the SEBS dielectric is completely destroyed by the solvent.


a
Crosslinked stretchable dielectric
Sacrificial layer (Dextran)
Si/SiO2
Stretchable semiconductor
Si/SiO2
1. Spin-coating of stretchable semiconductor film
Fluorinated sacrificial layer
2. Spin-coating of fluorinated polymer as sacrificial layer
Cu
3. Cu evaporation through shadow mask
4. Oxygen-plasma etching
5. Lifting-off of Cu by soaking in fluorinated solvent
b
CONPHINE film
R= C12H25 C10H21
x/y = 12:88
c

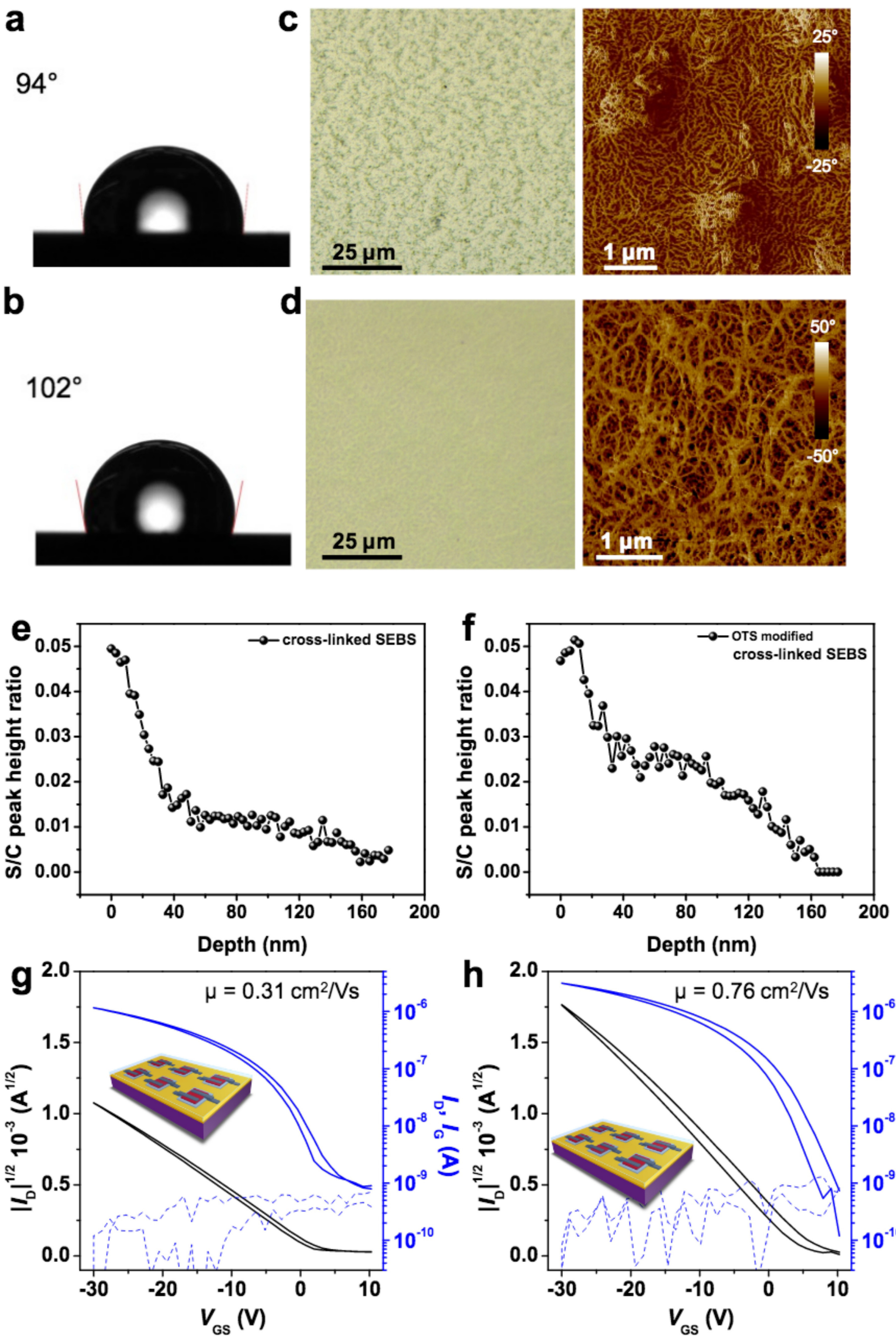
Extended Data Figure 3 | Etching-based patterning for stretchable semiconductors. **a**, Detailed patterning steps. **b**, Diagram of the CONPHINE semiconducting film, which is composed of 70 wt% SEBS (x/y = 12:88) and 30 wt% semiconductor phase, with both chemical structures shown on the right. **c**, Optical microscopic image of the patterned CONPHINE film aligned on top of SEBS patterns; scale bar, 700 μm.



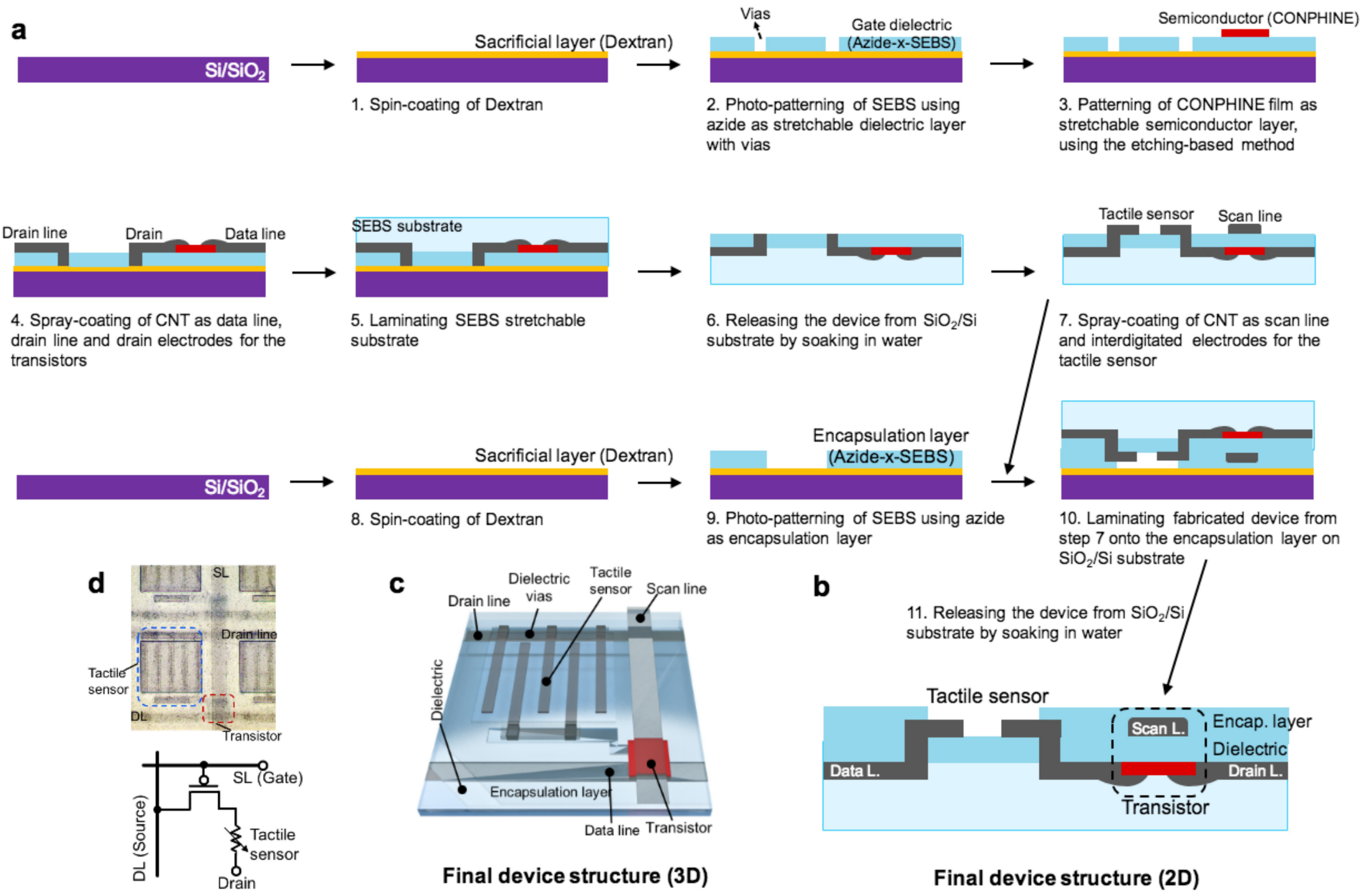
Extended Data Figure 4 | Test of the etching-based patterning method on different stretchable semiconductors. a, CONPHINE film. **b**, **c**, Its electrical performance in a thin-film transistor before (**b**) and after (**c**) patterning. **d**, A conjugated polymer (DPP2TTVT-PDCA) with hydrogen bonding that is built from 2,6-pyridine dicarboxamide (PDCA) moieties. **e, f**, Its electrical performance in in a thin-film transistor before (**e**) and after (**f**) patterning. All the thin-film transistors for testing have the device structure shown in Extended Data Fig. 2b.



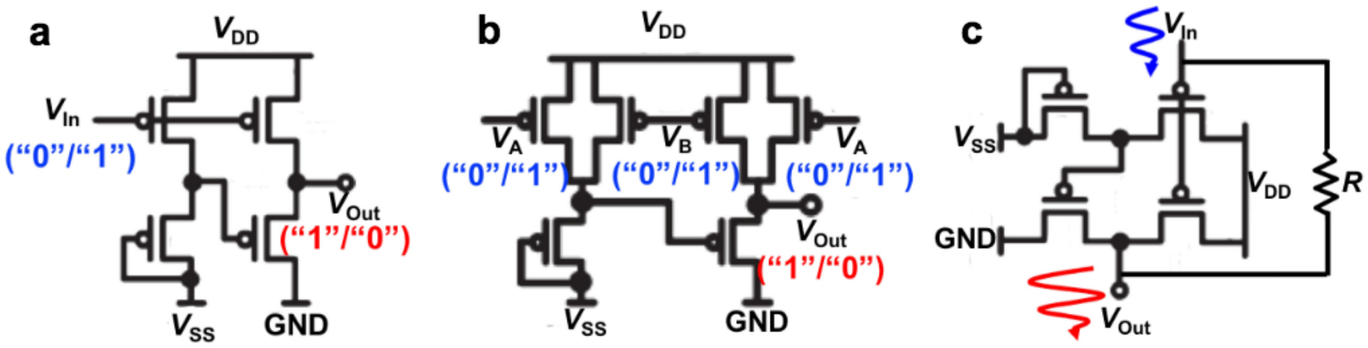
Extended Data Figure 5 | Inkjet-printed polymer semiconductor pattern on an azide-crosslinked SEBS dielectric. a, Optical microscope images of a typical printed pattern of 29-DPP-SVS-(2). **i**, Bright field. **ii**, Dark field. **iii**, **iv**, Cross-polarized images. **b**, Transfer characteristics of a transistor made from such an inkjet-printed semiconductor pattern, with thermally evaporated gold for the top contact electrodes. The transistor gives ideal transfer behaviour with no hysteresis. The obtained average mobility is 0.072 $cm^2 V^{-1} s^{-1}$. **c**, Corresponding output curve for this transistor. Here, even though its drop-casting-like deposition and crystallization of the semiconductor polymer may not give the best electrical performance (or, possibly, stretchability), the inkjet printing process is convenient as it involves fewer steps. **d**, Optical microscopic image of printed 29-DPP-SVS-(2) patterns aligned on top of SEBS patterns. Scale bar, 700 μm. **e**, Initial mobility and stretchability (the latter reflected by mobility at 100% strain along the charge-transport direction) of 29-DPP-SVS-(2) patterns obtained by inkjet printing, compared with patterns obtained by spin-coating with copper-protected etching. Inkjet printing gives slightly lower initial mobility than spin-coating, but slightly better stretchability.



Extended Data Figure 6 | Morphology and performance optimization of a CONPHINE film directly spin-coated on an azide-crosslinked SEBS surface, through surface modification with OTS. a, b, Contact angles of water on unmodified azide-crosslinked SEBS (**a**) and OTS-modified azide-crosslinked SEBS (**b**). **c, d**, Surface morphology (left, optical microscopic images; right, AFM images) of CONPHINE film on unmodified azide-crosslinked SEBS (**c**) and OTS-modified azide-crosslinked SEBS (**d**). **e**, **f**, XPS characterization of sulfur/carbon peak height ratios, showing the vertical distribution of the semiconductor phase in CONPHINE films obtained on unmodified azide-crosslinked SEBS (**e**) and on OTS-modified azide-crosslinked SEBS (**f**). Given that sulfur appears only in conjugated polymers, the ratio of the sulfur 2*p* to carbon 1*s* peaks qualitatively reflects the amount of conjugated polymer at different depths. **g**, **h**, Performance of transistor arrays fabricated on Si/SiO_2 substrates with CNTs as source/drain electrodes, spin-coated CONPHINE film as the semiconductor layer, and unmodified azide-crosslinked SEBS (**g**) or OTS-modified azide-crosslinked SEBS (**h**) as the dielectric layer.


a
Si/SiO$_2$
Sacrificial layer (Dextran)
1. Spin-coating of Dextran
Vias
Gate dielectric (Azide-x-SEBS)
2. Photo-patterning of SEBS using azide as stretchable dielectric layer with vias
Semiconductor (CONPHINE)
3. Patterning of CONPHINE film as stretchable semiconductor layer, using the etching-based method
Drain line
Drain
Data line
4. Spray-coating of CNT as data line, drain line and drain electrodes for the transistors
SEBS substrate
5. Laminating SEBS stretchable substrate
6. Releasing the device from SiO$_2$/Si substrate by soaking in water
Tactile sensor
Scan line
7. Spray-coating of CNT as scan line and interdigitated electrodes for the tactile sensor
Si/SiO$_2$
Sacrificial layer (Dextran)
8. Spin-coating of Dextran
Encapsulation layer (Azide-x-SEBS)
9. Photo-patterning of SEBS using azide as encapsulation layer
10. Laminating fabricated device from step 7 onto the encapsulation layer on SiO$_2$/Si substrate
11. Releasing the device from SiO$_2$/Si substrate by soaking in water
d
SL
Drain line
Tactile sensor
DL
Transistor
SL (Gate)
DL (Source)
Tactile sensor
Drain
c
Drain line
Dielectric vias
Tactile sensor
Scan line
Dielectric
Encapsulation layer
Data line
Transistor
Final device structure (3D)
b
Tactile sensor
Scan L.
Encap. layer
Dielectric
Data L.
Drain L.
Transistor
Final device structure (2D)

Extended Data Figure 7 | Fabrication and structure of an intrinsically stretchable, active-matrix tactile sensor array. a, Fabrication steps. **b**, Two-dimensional device structure. **c**, Three-dimensional device structure. **d**, Optical microscope image and circuit diagram of one pixel in the stretchable active-matrix tactile sensor array, showing the connection between the transistor and the corresponding tactile sensor. With the tactile sensors connected between the drain lines and the transistor drain electrodes through the dielectric's via holes, their resistance change from contact with conductive objects (including human skin) can be sampled by the drain currents through the transistors.



Extended Data Figure 8 | Circuit diagrams of the intrinsically stretchable circuits. a, Inverter. **b**, NAND. **c**, Amplifier.

Extended Data Table 1 | Comparison of our intrinsically stretchable transistor array with previous milestone works on stretchable transistors

		Buckling method		Rigid-island with stretchable interconnect		Intrinsically stretchable					
Ref.	**This work**	10	37	38	11	15	17	16	13	14	39
Semiconductor	29-DPP-SVS /SEBS *(polymer)*	DNTT (*organic small molecule*)	Si	Pentacene *(organic small molecule)*	DNTT (*organic small molecule*)	CNT	DPP2TTVT-PDCA *(polymer)*	DPPT-TT/SEBS *(polymer)*	CNT	P3HT *(polymer)*	CNT
Dielectric	SEBS	AlO_x/SAM	SiO_2	Polyimide	AlO_x/SAM	PU	PDMS	SEBS	PU	PU	$BaTiO_3$/PDMS
Conductor	CNT	Au	Cr/Au	Ag nanoparticles	Au	CNT	CNT	CNT	CNT	CNT, liquid metal	CNT
Mobility ($cm^2/(V·s)$)	1.37 (ave.) 1.78 (max.)	0.88 (ave.)	290 (n-type) 140 (p-type)	c.a. 0.48 (ave.)	1.8 for single transistor	27.0 (ave.) 32.5 (max.)	0.28 (ave.) 0.6 (max)	0.59 (ave)	0.18 (ave.)	0.034	~4
Density (cm^{-2})	347	c.a. 5	c.a. 280	*Not reported*	c.a. 4	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	c.a. 4 (4 devices)
Driving voltage (V)	10-30	5	5	100	3	8	60	40	60	80	30
On-current (μA/mm)	7.4	120	c.a. 1300	0.42	3.3	52	2.8	2.5	5	c.a. 0.73	14.4
Demonstrated stretchability	600% strain for 1 cycle; 100% strain for 1000 cycles	100% strain for 200 cycles (*single device*)	5% strain without cycling	70% strain without cycling	110% strain for 1 cycle (4 devices, density of ~0.4 cm^{-2})	50% strain for 1 cycle; 20% strain for 500 cycles	100% strain for 1 cycle; 25% strain for 500 cycles	100% strain for 1 cycle; 25% strain for 1000 cycles	100% strain for 1000 cycles	40% for 100 cycles	50% strain for 1400 cycles
Integrated functionality	Active matrix, digital and analog circuits	Active matrix (non-stretchable)	Digital and analog circuits	Active matrix	Active matrix (2 × 2)	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	Digital circuits (with non-ideal performance)

We compare data on our transistor array with data on previous milestone approaches to skin electronics[10,11,13–17,37–39], in terms of the materials used, key parameters on electrical performance, and skin conformability (represented by stretchability).

doi:10.1038/nature25754

Hadean silicate differentiation preserved by anomalous ^{142}Nd/^{144}Nd ratios in the Réunion hotspot source

Bradley J. Peters[1], Richard W. Carlson[1], James M. D. Day[2] & Mary F. Horan[1]

Active volcanic hotspots can tap into domains in Earth's deep interior that were formed more than two billion years ago[1,2]. High-precision data on variability in tungsten isotopes have shown that some of these domains resulted from differentiation events that occurred within the first fifty million years of Earth history[3,4]. However, it has not proved easy to resolve analogous variability in neodymium isotope compositions that would track regions of Earth's interior whose composition was established by events occurring within roughly the first five hundred million years of Earth history[5,6]. Here we report ^{142}Nd/^{144}Nd ratios for Réunion Island igneous rocks, some of which are resolvably either higher or lower than the ratios in modern upper-mantle domains. We also find that Réunion ^{142}Nd/^{144}Nd ratios correlate with helium-isotope ratios (^{3}He/^{4}He), suggesting parallel behaviour of these isotopic systems during very early silicate differentiation, perhaps as early as 4.39 billion years ago. The range of ^{142}Nd/^{144}Nd ratios in Réunion basalts is inconsistent with a single-stage differentiation process, and instead requires mixing of a conjugate melt and residue formed in at least one melting event during the Hadean eon, 4.56 billion to 4 billion years ago. Efficient post-Hadean mixing nearly erased the ancient, anomalous ^{142}Nd/^{144}Nd signatures, and produced the relatively homogeneous ^{143}Nd/^{144}Nd composition that is characteristic of Réunion basalts. Our results show that Réunion magmas tap into a particularly ancient, primitive source compared with other volcanic hotspots[7–10], offering insight into the formation and preservation of ancient heterogeneities in Earth's interior.

A remarkable feature of Earth's volcanic 'hotspots'—regions on the surface that may sit above plumes of magma rising from deep within the mantle—is that they can preserve geochemical signatures of ancient geological events, despite billions of years of mantle convection acting to erase these signatures. For example, ocean island basalts (OIBs) from Mangaia, the most southerly of the Cook Islands, preserve compelling geochemical evidence for the recycling of Archean atmospheric sulfur[1], and lavas from Hawai'i and Samoa may record metal–silicate separation that occurred in the early Earth[3]. The preservation of ancient geochemical signatures across multiple isotope systems implies that some geochemical domains may have remained relatively intact in Earth's deep interior for nearly all of its 4.56-billion-year history[3,4]. The strongest evidence for the preservation of early-formed compositional heterogeneity in Earth's deep interior is the recent detection of variability in the decay products of short-lived hafnium (^{182}Hf) and iodine (^{129}I) isotopes in modern mantle-derived rocks[3,4,11]. With half-lives of less than 20 million years (Myr), these two radioactive parents only track chemical differentiation events that occurred within the first 50–100 Myr of Earth history. The consequences of early Earth differentiation should be shown with even greater sensitivity by variability in ^{142}Nd, caused by the decay of samarium (^{146}Sm, with a half-life of 103 Myr; ref. 12), which would track events occurring in the first 500 Myr or so of Earth history. However, most previous efforts to detect variability in ^{142}Nd/^{144}Nd in modern volcanic hotspots and other mantle-derived rocks[5,6,13,14] have not yielded compelling evidence for the existence of variable ^{142}Nd/^{144}Nd ratios in the modern mantle. Two possible explanations for this observation are that isotopes of lithophile elements (those concentrated in Earth's crust, rather than core or mantle), such as samarium–neodymium, are susceptible to overprinting by recycled crustal material, or that total variation in ^{142}Nd/^{144}Nd amongst modern rocks is no greater than the analytical precision currently obtained for these isotope ratios.

Réunion Island, in the Indian Ocean, is an active volcanic hotspot that offers a unique perspective on ancient events preserved in the isotopic record of OIBs. Réunion OIBs have relatively homogeneous isotopic compositions that lie near the centre of the range displayed by other OIBs[8], implying a relatively pristine source that is distinct from the sources of other hotspots. This source has been suggested to have formed in the first 200 Myr of Earth's history[7] and contains limited recycled crustal material[9,10], meaning that it is ideal for studies investigating primordial Earth materials. Seismic studies have revealed large, sharply defined domains in Earth's deep mantle—including some beneath areas near Réunion—that are characterized by low or ultra-low shear-wave velocity[15] (termed 'large low-shear-velocity provinces', LLSVPs, and 'ultra-low velocity zones', ULVZs) and may be composed of material with a composition different to that of the rest of the mantle[16]. This material may represent sequestered primordial material[16], accumulated subducted oceanic crust[17], interaction of the mantle with Earth's core[18], sinking of dense silicate melts[19], or crystallization products of a basal magma ocean[20]. Advances in the understanding of LLSVPs and ULVZs through geochemistry may help to differentiate between these hypotheses. We show here that the Réunion mantle source possesses a statistically variable ^{142}Nd/^{144}Nd composition that is consistent with derivation from a mantle source that experienced a major Hadean silicate differentiation event. These new data may provide constraints on the ancient heritage of some LLSVPs.

We measured the neodymium isotopic compositions of 20 basalts and 2 cumulate xenoliths from Piton des Neiges and Piton de la Fournaise on Réunion (Fig. 1). We found that these samples are normally distributed about a mean of $\mu^{142}\text{Nd} = ((^{142}\text{Nd}/^{144}\text{Nd})_{\text{sample}}/(^{142}\text{Nd}/^{144}\text{Nd})_{\text{standard}} - 1) \times 10^{6} = +2.3 \pm 7.4$ (2 standard deviations, s.d.). Three replicated samples from this dataset are statistically resolvable from the terrestrial standard, JNdi, with μ^{142}Nd values of $+7.0 \pm 1.9$, $+6.3 \pm 1.7$ and -7.9 (range -5.3 to -9.0). A discussion of the relationship between JNdi and terrestrial geology is provided in the Methods. Reported precisions are means and 95% confidence intervals of all individual runs weighted by their respective internal precision or, where compiled data from multiple runs are slightly skewed, the *t*-test mean and 95% confidence interval of all compiled measurement cycles (typically more than 2,000 individual ratios). Statistical resolution from JNdi means that the Réunion source was at least partially influenced by a domain that differentiated from the rest of the mantle within the first 500 Myr or so of Earth history, while ^{146}Sm was still present. By contrast, Indian Ocean mid-ocean ridge basalt—which may

[1]Department of Terrestrial Magnetism, Carnegie Institution for Science, Washington DC 20015, USA. [2]Geosciences Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA.

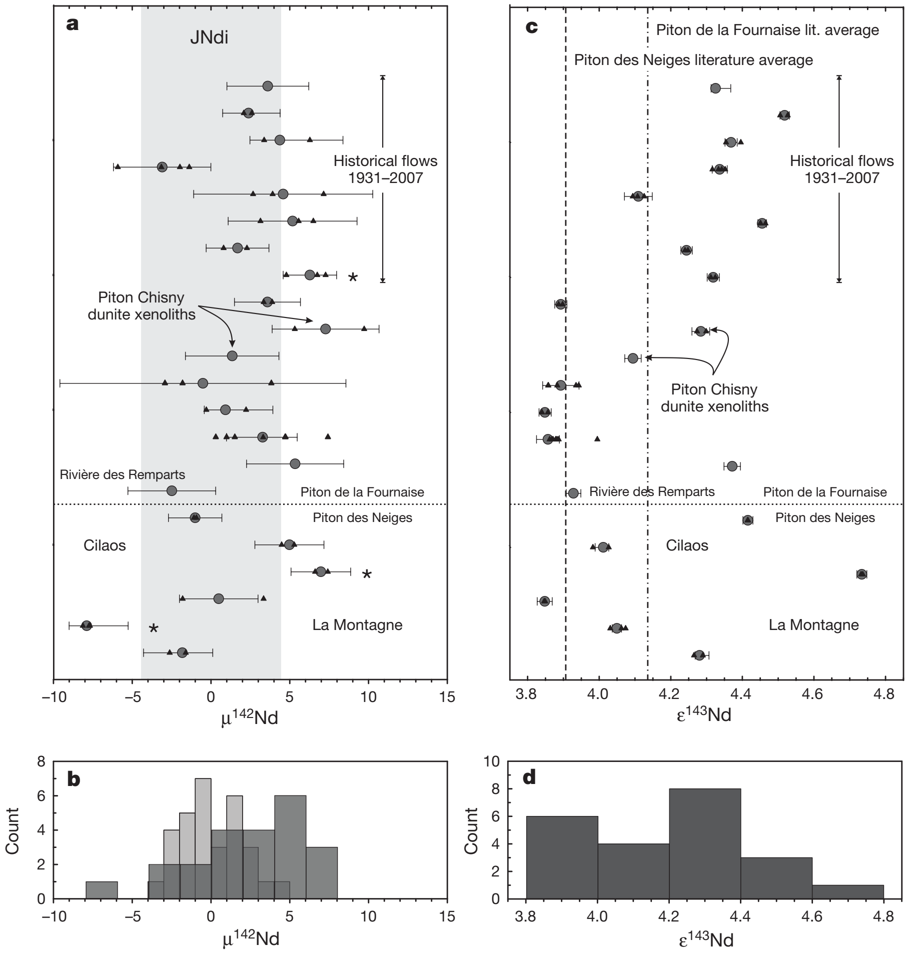


Figure 1 | Neodymium isotope data for Réunion OIB. We measured the Nd isotopic compositions of 20 basalts and 2 cumulate xenoliths from Réunion. We collected eight basalt samples from historical flows of the Piton de la Fournaise complex; eight basalt samples and both dunite samples (the latter from the Piton Chisny volcanic cone) from prehistoric flows of Piton de la Fournaise, including the flanks of Rivière des Remparts; and six basalt samples from the extinct Piton des Neiges complex, including samples from Cirque du Cilaos and La Montagne. **a**, Sample means (μ^{142}Nd; dark grey circles) and 95% confidence intervals, as well as individual runs for each sample (black triangles). The 2σ standard deviation of repeated measurements of the terrestrial standard, JNdi, is shown by the shaded region. Samples marked by asterisks are statistically resolved from JNdi. **b**, Histograms of the data from panel **a**, with the light grey bars reflecting data for JNdi and dark grey showing Réunion samples. **c**, Réunion ε^{143}Nd sample averages, 95% confidence intervals and individual runs. **d**, Histogram of data from panel **c**.

represent the regional depleted mantle—has been reported to have a μ^{142}Nd that is unresolvable from JNdi within 5 parts per million (p.p.m.)[5]. The total ^{143}Nd (ε^{143}Nd) on Réunion $((^{143}\text{Nd}/^{144}\text{Nd})_{\text{sample}}/(^{143}\text{Nd}/^{144}\text{Nd})_{\text{CHUR}} - 1) \times 10^4$, where 'CHUR' is the reference chondritic uniform reservoir value) is normally distributed about a mean of $+4.19 \pm 0.10$ (2 standard errors of the mean, s.e.m.; Fig. 1c, d). A correlation exists between μ^{142}Nd and ^{3}He/^{4}He, but not with ^{187}Os/^{188}Os (osmium) ratios (Fig. 2)—a striking attribute given analogous correlations between tungsten and helium isotopes in some other hotspot lavas[3].

The average μ^{142}Nd and ε^{143}Nd composition of Réunion samples could possibly be explained by a single-stage differentiation, 4.39 billion years ago, of a bulk Earth with a chondritic Sm/Nd ratio

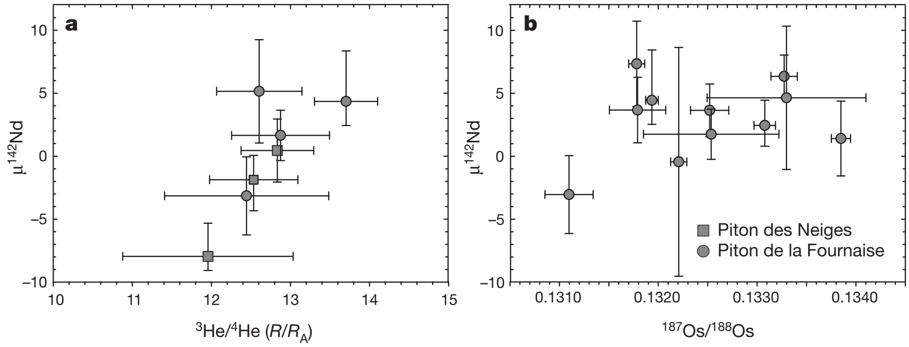


Figure 2 | Correlations between measured μ^{142}Nd values and other isotopic ratios from Réunion basalts. a, μ^{142}Nd versus ^{3}He/^{4}He (relative to the ratio of air, R_A; ref. 33). **b**, μ^{142}Nd versus ^{187}Os/^{188}Os (ref. 10). Error bars on helium and osmium isotope data reflect 2σ internal precision; error bars on μ^{142}Nd show 95% confidence intervals.

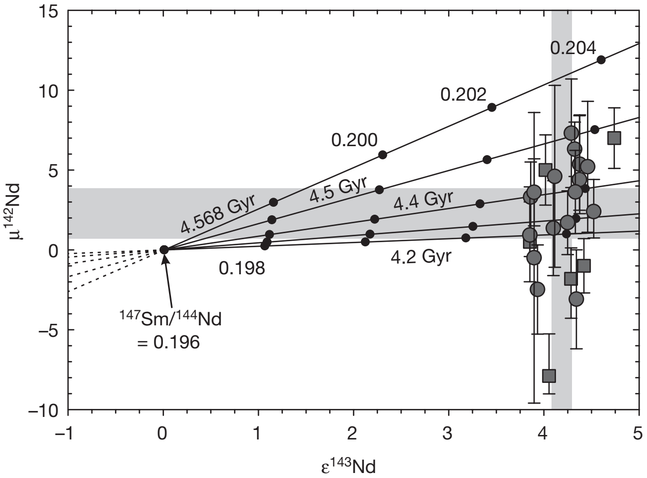


Figure 3 | μ^{142}Nd–ε^{143}Nd systematics of Réunion basalts and dunites, with curves representing single-stage evolution models. Model differentiation times are shown in billions of years (Gyr); other numbers within the graph are the ^{147}Sm/^{144}Nd ratios of differentiated reservoirs. Dark grey circles are the arithmetic means of sample ε^{143}Nd values (2 s.d.) and μ^{142}Nd values at 95% confidence interval. Réunion-average ε^{143}Nd (2 s.d.) and μ^{142}Nd (2 s.e.m.) values are shown with grey shading. Square symbols represent samples from Piton des Neiges and circles represent samples from Piton de la Fournaise, as in Fig. 2.

(^{147}Sm/^{144}Nd = 0.196) and a μ^{142}Nd value of 0 (see, for example, Fig. 3; see also Extended Data Fig. 1 for comparable scenarios with different bulk-Earth Sm/Nd and μ^{142}Nd ratios). This result is consistent with the observation that Réunion basalts lie on a lead–lead isochron formed around 4.39 billion years ago[7]. However, this relatively simple history for neodymium is unlikely for two reasons. First, closed-system evolution following a single-stage differentiation event would produce correlations between μ^{142}Nd and ε^{143}Nd, which are not seen in our data (Fig. 3a). Second, the observed total variability in μ^{142}Nd is greater than would be expected given external reproducibility about a singular average μ^{142}Nd–ε^{143}Nd (Extended Data Fig. 2). Consequently, we hypothesize that the Réunion source represents a binary mixture between an early depleted domain possessing a positive μ^{142}Nd, and an early enriched domain possessing a negative μ^{142}Nd that formed during the same differentiation event or during multiple, analogous differentiation events. The binary mixture may have formed in either the lower or the upper mantle, which could have implications for the formation of early magma oceans, or the formation of early crust.

We constructed a non-modal, accumulated fractional melting model for the lower-mantle mineral phases bridgmanite, calcium (Ca)-perovskite and ferropericlase (partitioning data from ref. 21). Our model minimizes misfit to Réunion μ^{142}Nd averages and extrema, average ε^{143}Nd and average ε^{176}Hf (ref. 8) ($\varepsilon^{176}\mathrm{Hf} = ((^{176}\mathrm{Hf}/^{177}\mathrm{Hf})_{\mathrm{sample}}/(^{176}\mathrm{Hf}/^{177}\mathrm{Hf})_{\mathrm{CHUR}} - 1) \times 10^4$) by varying the timing and degree of differentiation and the timing of mixing. In this model, assuming that the bulk Earth has a chondritic Sm/Nd ratio and a μ^{142}Nd of 0, a 78% partial melting event occurring 4.37 billion years ago would produce incompatible-element-enriched (melt) and -depleted (residue) reservoirs that would evolve to present-day values for μ^{142}Nd of −19 (melt) and +13 (residue), for ε^{143}Nd of −27 and +18, and for ε^{176}Hf of 34 and −1.5 (Fig. 4a, b) (incompatible elements are those that prefer the liquid state over the solid state). Subsequent mixing between these reservoirs forms a domain with a composition like that of Réunion (Fig. 4c). The proportion of each reservoir varies depending on the conditions of early differentiation, but typically involves 20–50% of the enriched (melt) component. Experimental constraints on lutetium–hafnium partitioning[21] in this scenario suggest a negative correlation between ε^{176}Hf and ε^{143}Nd that overlaps the isotopic composition of Réunion lavas[8] (Methods), which lie at the edge of the global 'mantle array' of Hf–Nd isotopes[22]. This predicted anti-correlation is at odds with the positive sense of the modern mantle Nd–Hf array for most recently erupted basalts, but may be required by the Nd–Hf isotopic characteristics of Archaean rocks[23,24] and has

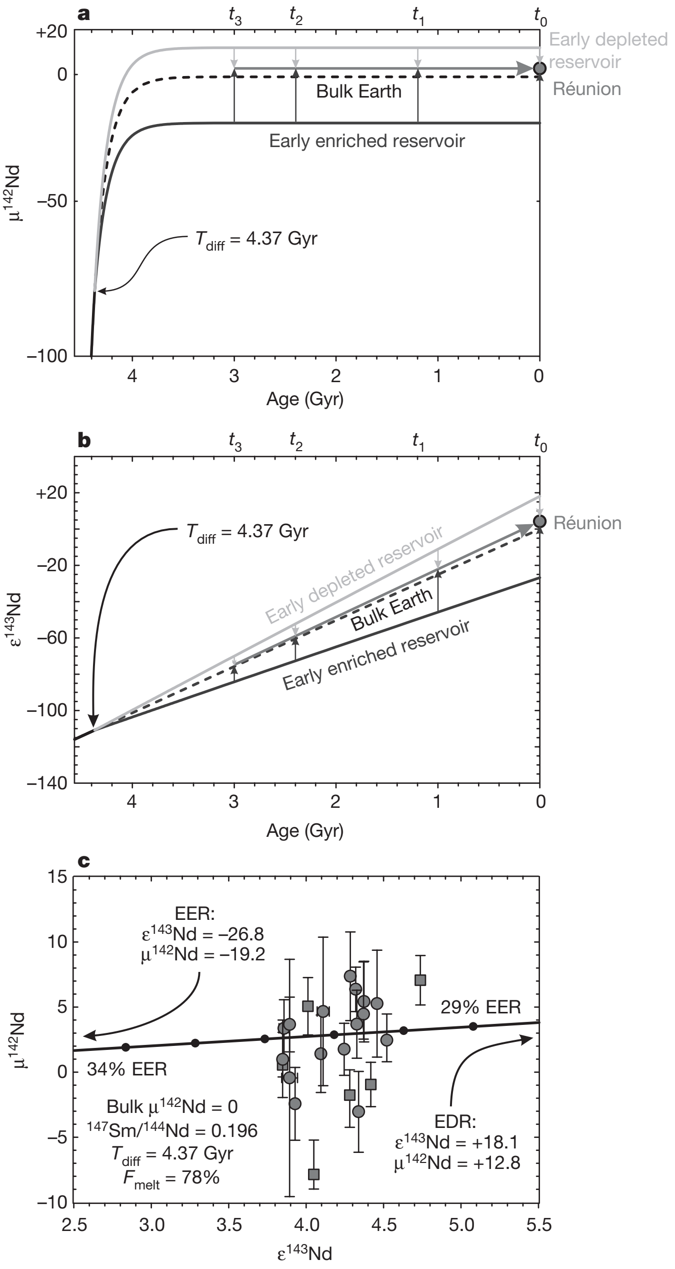


Figure 4 | Preferred model of differentiation 4.37 Gyr ago at uppermost lower-mantle pressures. Pressures less than 32–36 GPa; refs 26, 27. **a**, **b**, In our fractional melting model, differentiation of the initial Earth composition (T_{diff}) occurs at 4.37 Gyr ago; later, mixing occurs between conjugate melt (dark grey) and residue (light grey) at later times marked by vertical arrows to recreate the μ^{142}Nd (**a**) and ε^{143}Nd (**b**) compositions of the Réunion mantle source (medium grey lines). **c**, Réunion μ^{142}Nd–ε^{143}Nd compositions and 95% confidence interval (squares, Piton des Neiges; circles, Piton de la Fournaise) plotted with a mixing array between the conjugate melt and residue. Parallel mixing arrays with higher or lower μ^{142}Nd are possible if the Earth is assumed to have a different bulk μ^{142}Nd composition, and steeper mixing arrays are possible for different times and degrees of melting, but may also affect how well the model matches the average Réunion ε^{176}Hf composition (see text for details). Post-Hadean silicate differentiation may produce more homogenous ε^{143}Nd relative to μ^{142}Nd and, therefore, a steeper mixing array. EDR, early depleted reservoir; EER, early enriched reservoir.

been predicted by modelling of high-pressure mantle phases under conditions of magma ocean formation[25].

Using experimental constraints on samarium and neodymium partitioning in lower-mantle mineral phases[21], an idealized lower-mantle mineralogy, and melting sequences determined by diamond-anvil experiments[26,27], our model can reproduce the Réunion μ^{142}Nd signature given 50–90% partial melting in the deep Hadean mantle. The predicted degree of melting is strongly controlled by the rate of Ca-perovskite consumption during melting, as Ca-perovskite effectively buffers Sm/Nd fractionation during melting owing to its partitioning behaviour relative to bridgmanite (see Extended Data Figs 3–8 for an overview of various melting scenarios). Alternative scenarios that produce similar quantitative results are discussed in the Methods. Such high predicted degrees of melting could produce an iron-rich liquid if formed at high pressures[26]; this iron-rich liquid might have a higher density than surrounding perovskite mantle[28], and sink to form piles with LLSVP-like seismic properties[15]. Subsequent interaction between the melt and the proximate residuum, possibly as late as the beginning of plume ascent, could produce a mixed reservoir with geochemical characteristics similar to those observed for the Réunion hotspot (Fig. 4).

Our consideration of the deep mantle differentiation scenario is driven by the recognition that energetic events accompanying planetary accretion—for example, the giant impact that formed Earth's moon—probably instigated an episode of large-scale magmatism on Earth. Alternatively, the Hadean contributor to the Réunion mantle source could have differentiated at temperatures and pressures representative of the modern upper mantle. Garnet, a common upper-mantle phase, is known to substantially fractionate rare-earth elements, including samarium and neodymium, during melting. In a primitive silicate mantle possessing 10% modal garnet and a μ^{142}Nd of -10, accumulated, non-modal fractional melting of 1.4% (by mass) 4.26 billion years ago is sufficient to form a depleted mantle domain with μ^{142}Nd and Sm/Nd values similar to those of Earth's modern upper mantle ($\mu^{142\text{Nd}} \approx 0$; ^{147}Sm/^{144}Nd ≈ 0.25; partition coefficients from ref. 29). If melting locally progressed to 2.7%, this would form conjugate depleted and enriched domains with μ^{142}Nd values of $+20$ and -28 (Extended Data Fig. 9)—values that approach those of enriched Hadean crustal material[30]—and would reproduce the μ^{142}Nd mixing relationship shown in Fig. 4c. One challenge posed by an upper-mantle differentiation scenario, however, is that it may be more likely that distinct upper-mantle domains are completely remixed as vigorous upper-mantle convection is established between the Hadean and the present day. However, regardless of whether the Réunion source witnessed Hadean differentiation at low or high mantle pressures, its μ^{142}Nd signature suggests that it may be related to foundational Earth events.

There are few constraints on helium solubility during mantle differentiation, but helium-solubility experiments suggest that melting residua may be enriched in helium relative to uranium and thorium[31]. If this is the case, then higher ^{3}He/^{4}He ratios would be associated with positive μ^{142}Nd values, and lower ^{3}He/^{4}He ratios would be associated with negative μ^{142}Nd values, as we observe in the Reunion data (Fig. 2a). Mixing between these two reservoirs would result in a positive correlation between μ^{142}Nd and ^{3}He/^{4}He that may survive overprinting in the partially degassed upper mantle. The ^{187}Os/^{188}Os signature of Réunion lavas may be related to the same mixing event if one reservoir effectively overprinted the other. This would preclude a statistically significant correlation between μ^{142}Nd and ^{187}Os/^{188}Os (Fig. 2b), similar to the case that we argue for μ^{142}Nd and ε^{143}Nd (Fig. 3a).

The presence of LLSVPs in Earth's deep mantle is remarkable given the dynamic behaviour that characterizes the upper mantle and mantle transition zone. The existence of mantle LLSVPs has been suggested to reflect continuous delivery of subducted material to the deep mantle[17]; however, the petrology[32] and geochemistry[9] of Réunion lavas do not suggest a major contribution from recycled crust. On the other hand, a major Hadean deep-mantle differentiation could have produced a dense, enriched domain and a geochemically complementary depleted domain that persisted in close proximity for billions of years. The nature of this event cannot easily be determined, but given the high degrees of melting required by perovskite partitioning data, it may be related to the formation of a magma ocean within the energetic early Earth. Our results show that the Réunion hotspot source preserves a geochemical vestige of an early silicate differentiation event that might be related to LLSVP formation in the deep mantle. This provides strong evidence that primordial silicate domains can survive in the deep Earth for billions of years, despite sustained mantle convection. Réunion remains a major untapped source of information about early Earth processes that complements research on the behaviour of short-lived radiogenic isotope systems in other global hotspots.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 September 2017; accepted 2 January 2018.

1. Cabral, R. A., Jackson, M. G., Rose-Koga, E. F., Koga, K. T., Whitehouse, M. J., Antonelli, M. A., Farquhar, J., Day, J. M. D. & Hauri, E. H. Anomalous sulphur isotopes in plume lavas reveal deep mantle storage of Archaean crust. *Nature* **496,** 490–493 (2013).
2. Delavault, H., Chauvel, C., Thomassot, E., Devey, C. W. & Dazas, B. Sulfur and lead isotopic evidence of relic Archean sediments in the Pitcairn mantle plume. *Proc. Natl Acad. Sci. USA* **113,** 12952–12956 (2017).
3. Mundl, A. *et al.* Tungsten-182 heterogeneity in modern ocean island basalts. *Science* **356,** 66–69 (2017).
4. Rizo, H. *et al.* Preservation of Earth-forming events in the tungsten isotopic composition of modern flood basalts. *Science* **352,** 809–812 (2016).
5. Jackson, M. G. & Carlson, R. W. Homogenous superchondritic ^{142}Nd/^{144}Nd in the mid-ocean ridge basalt and ocean island basalt mantle. *Geochem. Geophys. Geosyst.* **13,** (2012).
6. Horan, M. F., Carlson, R. W., Walker, R. J., Jackson, M. & Garçon, M. Tracking Hadean processes in modern basalts. *Earth Planet. Sci. Lett.* **484,** 184–191 (2018).
7. Vlastélic, I., Lewen, E. & Staudacher, T. Th/U and other geochemical evidence for the Réunion plume sampling a less differentiated mantle domain. *Earth Planet. Sci. Lett.* **248,** 379–393 (2006).
8. Bosch, D. *et al.* Pb, Hf and Nd isotope compositions of the two Réunion volcanoes (Indian Ocean): a tale of two small-scale mantle "blobs"? *Earth Planet. Sci. Lett.* **265,** 748–765 (2008).
9. Schiano, P. *et al.* Osmium isotopic systematics of historical lavas from Piton de la Fournaise (Réunion Island, Indian Ocean). *Contrib. Mineral. Petrol.* **164,** 805–820 (2012).
10. Peters, B. J., Day, J. M. D. & Taylor, L. A. Early mantle heterogeneities in the Réunion hotspot source inferred from highly siderophile elements in cumulate xenoliths. *Earth Planet. Sci. Lett.* **448,** 150–160 (2016).
11. Mukhopadhyay, S. Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* **486,** 101–104 (2012).
12. Friedman, A. M. *et al.* Alpha decay half lives of ^{148}Gd, ^{150}Gd and ^{146}Sm. *Radiochim. Acta* **5,** (1966).
13. Caro, G., Bourdon, B., Halliday, A. N. & Quitté, G. Super-chondritic Sm/Nd ratios in Mars, the Earth and the Moon. *Nature* **452,** 336–339 (2008).
14. Murphy, D. T., Brandon, A. D., Debaille, V., Burgess, R. & Ballentine, C. In search of a hidden long-term isolated sub-chondritic ^{142}Nd/^{144}Nd reservoir in the deep mantle: implications for the Nd isotope systematics of the Earth. *Geochim. Cosmochim. Acta* **74,** 738–750 (2010).
15. Garnero, E. J. & McNamara, A. K. Structure and dynamics of Earth's lower mantle. *Science* **320,** 626–628 (2008).
16. Wen, L., Silver, P., James, D. & Kuehnel, R. Seismic evidence for a thermo-chemical boundary at the base of Earth's mantle. *Earth Planet. Sci. Lett.* **189,** 141–153 (2001).
17. Mulyukova, E., Steinberger, B., Dabrowski, M. & Sobolev, S. V. Survival of LLSVPs for billions of years in a vigorously convecting mantle: replenishment and destruction of chemical anomaly. *J. Geophys. Res.* **120,** 3824–3847 (2015).
18. Dubrovinsky, L. *et al.* Iron–silica interaction at extreme conditions and the electrically conducting layer at the base of Earth's mantle. *Nature* **422,** 58–61 (2003).
19. Lee, C. T. A. *et al.* Upside-down differentiation and generation of a 'primordial' lower mantle. *Nature* **463,** 930–933 (2010).
20. Labrosse, S., Hernlund, J. W. & Coltice, N. A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature* **450,** 866–869 (2007).
21. Corgne, A., Liebske, C., Wood, B. J., Rubie, D. C. & Frost, D. J. Silicate perovskite-melt partitioning of trace elements and geochemical signature of a deep perovskitic reservoir. *Geochim. Cosmochim. Acta* **69,** 485–496 (2005).
22. Chauvel, C., Lewin, E., Carpentier, M., Arndt, N. T. & Marini, J.-C. Role of recycled oceanic basalt and sediment in generating the Hf–Nd mantle array. *Nat. Geosci.* **1,** 64–67 (2008).

23. Blichert-Toft, J., Arndt, N. T., Wilson, A. & Coetzee, G. Hf and Nd isotope systematics of early Archean komatiites from surface sampling and ICDP drilling in the Barberton Greenstone Belt, South Africa. *Am. Mineral.* **100,** 2396–2411 (2015).
24. Rizo, H., Boyet, M., Blichert-Toft, J. & Rosing, M. Combined Nd and Hf isotope evidence for deep-seated source of Isua lavas. *Earth Planet. Sci. Lett.* **312,** 267–279 (2011).
25. Caro, G., Bourdon, B., Wood, B. J. & Corgne, A. Trace-element fractionation in Hadean mantle generated by melt segregation from a magma ocean. *Nature* **436,** 246–249 (2005).
26. Tateno, S., Hirose, K. & Ohishi, Y. Melting experiments and peridotite to lowermost mantle conditions. *J. Geophys. Res. Solid Earth* **119,** 4684–4694 (2014).
27. Fiquet, G. *et al.* Melting of peridotite to 140 gigapascals. *Science* **329,** 1516–1518 (2010).
28. Funamori, N. & Sato, T. Density contrast between silicate melts and crystals in the deep mantle: an integrated view based on static-compression data. *Earth Planet. Sci. Lett.* **295,** 435–440 (2010).
29. Salters, V. & Longhi, J. Trace element partitioning during the initial stages of melting beneath mid-ocean ridges. *Earth Planet. Sci. Lett.* **166,** 15–30 (1999).
30. O'Neil, J., Carlson, R. W., Francis, D. & Stevenson, R. K. Neodymium-142 evidence for Hadean mafic crust. *Science* **321,** 1828–1831 (2008).
31. Parman, S. W., Kurz, M. D., Hart, S. R. & Grove, T. L. Helium solubility in olivine and implications for high ^{3}He/^{4}He in ocean island basalts. *Nature* **437,** 1140–1143 (2005).
32. Sobolev, A. V., Hofmann, A. W., Sobolev, S. V. & Nikogosian, I. K. An olivine-free mantle source of Hawaiian shield basalts. *Nature* **434,** 590–597 (2005).
33. Füri, E. *et al.* Helium isotope variations between Réunion Island and the Central Indian Ridge (17°–21°S): new evidence for ridge-hot spot interaction. *J. Geophys. Res.* **116,** B02207 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements B.J.P. is grateful for conversations with fellow postdocs A. Mundl and J. Reimink. Technical assistance with Nd isotope measurements was provided by T. Mock, and data-reduction routines were modified from an original version by M. Garçon. Funding for fieldwork for this study was provided by the National Geographic Society (NGS 8330-07), the Geological Society of America (GSA 10539-14), and a personal donation from R. Rex. Support for laboratory work was provided by the Carnegie Institution for Science.

Author Contributions B.J.P. performed chemical separations and isotope analyses and wrote the manuscript. R.W.C. and B.J.P. interpreted the data and developed the modelling. J.M.D.D. and B.J.P. collected rock samples. M.F.H. developed the chemical separation method and contributed to data analysis. All authors reviewed, edited and approved the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to B.J.P. (bpeters@carnegiescience.edu).

Reviewer Information *Nature* thanks B. Bourdon and B. Hanan for their contribution to the peer review of this work.

METHODS

Sample preparation and measurement of ^{142}Nd/^{144}Nd data. Twenty basalt and two dunite samples from Réunion Island were selected for this study. Of these, eight basalts were collected from historical flows of the Piton de la Fournaise complex; eight basalts and both dunites were collected from prehistoric flows of Piton de la Fournaise; and six basalts were collected from the extinct Piton des Neiges complex. The dunites are cumulate xenoliths from the Piton Chisny volcanic cone, located around 2 km east of the western rim of the Enclos Fouqué (the caldera). Data on major, trace and highly siderophile elements from these samples are reported in ref. 10, as are ^{187}Os/^{188}Os compositions. Helium isotopic compositions are reported in ref. 33 for select samples.

We carried out all chemical separations in the clean laboratory of the Department of Terrestrial Magnetism (DTM) at the Carnegie Institution for Science. Aliquots of 0.05–0.15 g of homogenized bulk-rock powder were digested in a solution comprising a 4/1 ratio of ultrapure hydrofluoric acid (HF) and Teflon-distilled nitric acid (HNO_3) at more than 120 °C for three days in closed beakers on hotplates, then evaporated to incipient dryness. Following this, they were taken up in around 1 ml HNO_3, capped and allowed to equilibrate at about 100 °C for at least 1 hour. Sample mixtures were again evaporated to incipient dryness, then taken up again in about 4 ml HNO_3 and digested at more than 120 °C for 24 hours. Resulting sample mixtures were evaporated to incipient dryness, then taken up in about 1 ml of Teflon-distilled hydrochloric acid (HCl), capped and allowed to equilibrate at around 100 °C for at least 1 hour. These sample mixtures were again evaporated to incipient dryness, then taken up again in about 4 ml HCl and digested at more than 120 °C for 24 hours. In nearly all cases, this procedure produced clear solutions. However, for dunite sample CH0703B and one aliquot of CH0703 Dunite (separate rocks collected at the same sampling location), the HF/HNO_3 digestion was repeated in Parr bombs in order to digest refractory phases. All samples were processed through a four-stage column separation procedure as follows.

First, samples were taken up in 2 ml 2 M HCl per 100 mg of digested rock and passed through a column containing 2 ml of a cation-exchange resin (AG50W-X8, which has a mesh size of 200–400). Major elements and heavy rare-earth elements (REEs) were removed from the column with sequential elutions of 2 M, 2.5 M and 6 M HCl; light REEs were then eluted in 6 M HCl. Following this, samples were taken up in 10 M HNO_3 plus 20 mM sodium bromate, and twice passed through a column containing 2 ml of 100–200 μm LN-Spec resin to remove most of the cerium (Ce) in the sample. Next, samples were taken up in 2.5 M HCl and again passed through a column containing 2 ml AG50W-X8 to remove residual sodium from the previous column. Finally, samples were taken up in 0.172–0.202 M HCl and passed through a long-aspect ratio column containing about 3 ml of 20–40 μm LN-Spec resin. Neodymium was chromatographically separated from Ce and Sm on this column using specific volumes of 0.172–0.202 M HCl calibrated for each acid batch and column. Total yields were typically 60–90% and total yields were most sensitive to yields from the second and fourth columns. Two total procedural blank analyses were performed, yielding 66 pg and 70 pg Nd, or less than 0.01% of a typical total sample Nd load. For a detailed, calibrated description of the separation method, see ref. 6.

We measured high-precision ^{142}Nd/^{144}Nd and ^{143}Nd/^{144}Nd data in positive-ion mode on a Thermo Fisher Scientific Triton thermal ionization mass spectrometer at DTM. The typical Nd load was 600–750 ng. The analysis method used nine Faraday cups in a four-step multidynamic routine, with ^{143}Nd, ^{144}Nd, ^{145}Nd and ^{146}Nd in the centre cup. Individual runs typically consisted of 600–900 cycles of four eight-second integrations; each cycle can produce two dynamic-corrected ^{142}Nd/^{144}Nd measurements. A detailed explanation of this method is available in ref. 35. A typical ^{142}Nd signal was 3–4 V, although samples with relatively little Nd were run with a slightly lower signal (greater than 2 V). Data were fractionation corrected to ^{146}Nd/^{144}Nd = 0.7219 using the exponential law. The ^{146}Nd/^{144}Nd value used for mass fractionation correction was interpolated to the same time of ^{142}Nd/^{144}Nd integration using the two nearest ^{146}Nd/^{144}Nd integrations in the same cups used to measure ^{142}Nd/^{144}Nd. Using this interpolated correction substantially improved data reproducibility over the standard method, which uses ^{146}Nd/^{144}Nd measured in the same cups as ^{142}Nd/^{144}Nd, but in the next mass step. When the ^{146}Nd/^{144}Nd rapidly changed rate or direction during a run, and when exponential-law correction could not reconcile measured ^{142}Nd/^{144}Nd to within five standard errors of the whole-run average over a cumulative average of fifty integrations or more, then we removed measurement cycles from the data correction until one of these conditions was no longer true. Sample ^{142}Nd/^{144}Nd and ^{143}Nd/^{144}Nd data are reported in Supplementary Tables 1, 2. The data are also available from the EarthChem database[36].

External precision was monitored using 31 measurements of the JNdi Nd standard over the course of the analysis campaign. During this time, the standard was re-diluted once and the Faraday cups were replaced. As a result, average measured values of JNdi for each condition (before re-dilution/before cup replacement, after re-dilution/before cup replacement, and after re-dilution/after cup replacement) were corrected to the average JNdi measurement after re-dilution, but before cup replacement, which represents the largest number of measurements under a single condition (n = 13). The 2 s.d. range of ^{142}Nd/^{144}Nd for all standards after this correction is 4.4 p.p.m., and we report this as our external precision. All JNdi ^{142}Nd/^{144}Nd values are given in Supplementary Table 2 and Extended Data Fig. 10.

Use of statistics in reporting ^{142}Nd/^{144}Nd results. All runs reported in Supplementary Table 1 are reported as averages and 95% confidence intervals of all dynamic-corrected measurements included in the dataset for the sample. To produce Figs 1–3, we compiled all runs of a given sample and calculated their arithmetic mean and 95% confidence interval. The calculations, which were performed in Isoplot, give greater weight to runs with greater precision. Many replicate runs represent multiple aliquots of a single digestate—that is, the same material is being measured twice and ought to be treated as a single 'true' composition represented within the precision of the measurements (note that all replicates of single digestates agree within error). However, we argue this similar treatment is faithful to data acquisition even when representing multiple digestates because, first, powder aliquots were taken from a homogenized batch originally at least 100 times the mass of the aliquots, and second, unlike other isotope systems, Nd is not concentrated in single-phase 'nuggets' that may bias data between powder aliquots. In many cases, the compiled data were normally distributed and the 95% confidence interval approximates the 2 s.e.m. for the same data.

In some cases, the compiled data for each sample were slightly skewed, and in these cases the 95% confidence interval is calculated in SigmaPlot using a Wilcoxon *t*-test. In these cases, the confidence intervals appear asymmetrical about the arithmetic mean and are larger than 2 s.e.m. in the direction of the skew longness. Skewness often arose owing to a single run that displayed a rapid or anomalous fractionation pattern that was compiled alongside runs that did not show such fractionation. Such a condition may arise, for example, when analysing the last aliquot of a sample and the Nd signal decreases rapidly before the intended end of a run. Although some of these data are removed from analysis according to the criteria mentioned above, skewness can be introduced by the remaining data and may have relevance to the 'true' ^{142}Nd/^{144}Nd value of a given sample. Accepting the ratios that introduce skew provides a more conservative view of the range of possible ^{142}Nd/^{144}Nd values for each sample, *ceteris paribus*, than a 95% confidence interval that approaches the 2σ error for the same sample without the skewed ratios. We therefore elect to present the data in this manner.

Occasionally, the confidence interval does not encompass the means of all individual runs, which may arise because of two circumstances. First, the confidence interval reflects the clustering of hundreds of measured ratios about the arithmetic mean of all runs, which are compiled together. This means that measured ratios from some runs that lie above or below the arithmetic mean for that run are stacked with other ratios from other runs. The arithmetic mean of all runs lies at the centre of a distribution that encompasses measured ratios that vary by 200–300 p.p.m., meaning that these differences are small relative to all compiled ratios. Second, the confidence interval may lie far from early runs, which tended to have slightly higher μ^{142}Nd values than later runs of the same sample. The reason for this is unclear, and there is no sound scientific or statistical argument known to us that would allow us to exclude these analyses entirely; however, in all cases these runs were shorter than nearly all subsequent runs, meaning that the weighted arithmetic mean and standard error are biased away from these aberrant values when the runs are compiled together, producing an unrealistically small standard error. In these instances (BPC1515 and BPC1517), the 95% confidence interval provides a more conservative view of our knowledge of 'actual' sample compositions.

By contrast, we use a standard-deviation approach to characterize standard data. JNdi is a 'terrestrial' standard, although its geologic origin is not precisely known. As a result, the isotopic composition of JNdi is primarily used to evaluate instrumental measurement reproducibility and to correct for interlaboratory biases, if they exist. The ^{143}Nd/^{144}Nd of JNdi is substantially lower than the chondritic value, and there is no clear reason to expect that JNdi was intentionally created to match the composition of the bulk silicate Earth. However, limited measurements of upper-mantle-derived rocks[5,6,13,14] indicate that the depleted mantle may lie within some 5 p.p.m. of the long-term JNdi average ^{142}Nd/^{144}Nd composition. Recent advances in knowledge of the stable isotopic composition of Earth strongly suggest that it does not have a μ^{142}Nd composition as markedly negative as that of carbonaceous chondrites[37], and may have a μ^{142}Nd closer to zero[38]. However, total μ^{142}Nd variability in terrestrial rocks is much greater (38.5 p.p.m.) than that expected in the upper mantle (5–10 p.p.m.) or due to uncertainty in the nucleosynthetic composition of Earth (around 6 p.p.m.). Consequently, we evaluate JNdi relative to the reservoir(s) it is intended to represent, rather than a statistical evaluation of our analytical prowess. This treatment would yield a 'precision' similar in magnitude to our 2 s.d. external precision (here 4.4 p.p.m., but in the recent past closer to 5 p.p.m.), which is the reporting norm at present and the one we favour here.

We determine whether a run or sample is resolved from the JNdi terrestrial standard by comparing the precision of a run or sample to the 2 s.d. of all standard runs. For example, if a run has a minimum permitted value of +4.8 p.p.m. based on a run mean of +8.5 p.p.m. and an internal precision (reported as 2 standard errors of the run mean) of 3.7 p.p.m., then we consider that run to be 'resolved' from the standard. Typically, internal precisions on individual standard and sample runs were indistinguishable and averaged 2–4 p.p.m. for runs of at least 600 cycles (1,200 ratios). Although individual runs from six samples (CH0703 Dunite, RU0709, RU0711, RU0716, RU0719, RU1515 and RU1517) are resolved from the standard, the means and 95% confidence intervals of only three samples (RU0709, RU0711 and RU0716) are resolved from JNdi. Both runs of sample CH0703 Dunite have elevated means (+9.8 p.p.m. and +5.3 p.p.m.) but poor precision owing to the relatively small abundance of Nd in this sample. As a result, the compiled runs are not resolvable from JNdi. Sample RU0709 is only slightly resolved owing to consistency between replicate measurements with modestly elevated means; however, each individual run is not resolved from JNdi. Samples RU0711 and RU0716 are strongly resolved from JNdi across multiple analyses in opposite directions. Strictly interpreted, these results are inconsistent with the notion that the Réunion source exclusively includes components that have a composition similar to the depleted upper mantle or continental crust.

Isotope modelling. Our Réunion μ^{142}Nd and ε^{143}Nd data are normally distributed, which is remarkable given that the sources of many hotspots represent dynamic mixtures of mantle domains with distinct isotopic means. The statistical robustness of the normality might mean that our measurements predominantly represent a well-mixed mantle source of singular μ^{142}Nd composition, which is similar to what is thought of the ^{87}Sr–^{143}Nd–^{206}Pb composition of Réunion basalts[7,8]. Because of this broad agreement between geological and statistical arguments, we would argue that it is valid to represent the Réunion Island μ^{142}Nd and ε^{143}Nd composition as an average ± 2 s.e.m. (for Réunion, +2.6 ± 1.6; $n = 22$). If this is a true interpretation of the total data variability, one might expect that the Réunion source was produced in a single differentiation event that produced a reservoir that is slightly depleted in incompatible elements (that is, it has an elevated and constant Sm/Nd relative to the bulk silicate Earth). This scenario can adequately represent the Réunion data for a variety of bulk-Earth μ^{142}Nd and Sm/Nd ratios (Extended Data Fig. 1). However, if the bulk Earth possesses μ^{142}Nd values of less than −14 and the Réunion source was derived from the bulk silicate Earth, then the resulting best-fit μ^{142}Nd–ε^{143}Nd for the Réunion source may be too unradiogenic to be supported by our data precision. To evaluate scenarios in Extended Data Fig. 1, we set a maximum best-fit misfit threshold of 21%, which is the ratio of our 4.4 p.p.m. two standard deviations on JNdi analyses to the total range of ^{142}Nd values of 20.8 p.p.m. measured across all individual runs. The misfit is calculated as a relative deviation from the mean μ^{142}Nd, ε^{143}Nd and ε^{176}Hf values of Réunion igneous rocks.

Notwithstanding, the standard deviation (as opposed to standard error discussed before) of the mean for all Réunion basalts (+2.3 ± 7.4) measured here is much larger than the standard deviation of JNdi measurements (0 ± 4.4) (that is, the observed normal distribution is much 'fatter' than the distribution of JNdi measurements), which strongly implies that the total variability in Réunion μ^{142}Nd is greater than would be expected from analytical precision alone, and, correspondingly, than what would be expected to arise as a result of a single-stage differentiation scenario.

To test this idea, we use a statistical hypothesis testing procedure similar to 'bootstrapping'. We create a synthetic dataset by generating a group of random numbers equal in size to the number of samples measured ($n = 22$) with an identical mean and a standard deviation that is set to the standard deviation of JNdi measurements. We then create a resampled dataset of our own data by randomly choosing, with replacement, a number of data points equal to the number of samples measured. The mean and standard deviation of these two datasets is calculated, and this procedure is repeated 500 times (see Source Data for an example). The result is shown in Extended Data Fig. 2, in which the standard error (which here is a constant function of standard deviation at fixed n) of the resampled data is larger than the standard error of the synthetic dataset (total overlap is less than 10%). Interestingly, there is a weak anticorrelation between the mean and standard error of the resampled dataset, which may imply that the 'true' average μ^{142}Nd of Réunion lavas is slightly greater than is implied by our sample set.

Importantly, this statistical analysis highlights the need to measure a large sample set to determine total μ^{142}Nd variability in a single hotspot. Given that many samples possess μ^{142}Nd values close to the hotspot mean, but that only three are statistically resolvable from JNdi, it might be reasonable to expect that a similar hotspot tapping a reservoir with a μ^{142}Nd that is statistically resolvable from JNdi will only have resolvable μ^{142}Nd values in around 15% of samples (for example, 3 samples out of 22, as observed for Réunion, is approximately a 14% 'success rate'). A similar geological argument can be made in general for μ^{142}Nd in the Réunion source, which may primarily reflect the effects of vigorous post-Hadean mixing typical of ascending mantle plumes, and therefore might have produced many lavas with 'normal' μ^{142}Nd composition and only a few that preserve the 'anomalous' μ^{142}Nd composition of their source. In this case, the exact μ^{142}Nd composition of the bulk silicate Earth is less important than the total range of sample μ^{142}Nd values. If samples preserve a total μ^{142}Nd range that is greater than present analytical precision, they must preserve different views of Hadean differentiation. The same is probably not true of long-lived radiogenic isotope systems, such as ^{147}Sm–^{143}Nd, where total variability within a single hotspot is much greater than analytical precision, or even for tungsten-isotope composition, which may be less likely to be overprinted by the silicate mantle. The composition of long-lived radiogenic isotope systems (such as ^{147}Sm–^{143}Nd, measured as ε^{143}Nd), however, may reflect the conditions of the last mixing or differentiation event, which might have occurred well after ^{146}Sm had decayed away.

A similar 'success rate' was observed for helium-isotope measurements of young basalts from the Rungwe province of East Africa[39], where only some 25% of measured samples have ^{3}He/^{4}He ratios that are resolvably greater than those of the upper mantle. Although this result may primarily be due to degassing, Nd isotopic compositions are highly susceptible to overprinting in the silicate mantle, which could be a process of equal pervasiveness. This might be an additional qualitative standpoint from which to evaluate the relationship between μ^{142}Nd and ^{3}He/^{4}He: although the Réunion source appears to lack extremely high ^{3}He/^{4}He ratios (which are present at other hotspots), or highly radiogenic μ^{142}Nd (which is present in some Archaean crustal rocks[40]), the correlation between these two isotope systems strongly implies that the Réunion source composition was formed from a domain with a He–Nd isotope composition even more extreme than that of any sample yet measured. A robust diagnosis of endmember compositions in either isotope system for a single setting may, in other words, require analysis of a large number of samples. So far, however, this extrapolation does not strictly match what is known of μ^{142}Nd in other settings of intraplate magmatism, such as the high-^{3}He/^{4}He Baffin Island picrites[6].

Another hypothesis is that the normal distribution of data is created by a relatively constant degree of mixing between depleted and enriched reservoirs. These reservoirs need not be mutually complementary (that is, formed in the same event); however, we evaluate the hypothesis under this assumption as the simplest scenario. We developed a numerical model designed to evaluate melting scenarios under deep mantle conditions (that is, at pressures of more than 23 GPa), followed by mixing between conjugate enriched and depleted reservoirs, assuming closed-system evolution before and after the mixing event. We assume mantle mineral modes of 81% bridgmanite, 9% Ca-perovskite and 10% ferropericlase, and use the bridgmanite and Ca-perovskite partition coefficients for Nd, Sm, Lu and Hf from ref. 21. We assume that ferropericlase partitioning of these elements is 0, meaning that the ferropericlase melting buffers Sm/Nd and Lu/Hf fractionation. We assume that all phases melt on the solidus, although this is difficult to confirm in high-pressure experiments. We then set the order in which the three phases are consumed on the basis of three endmember scenarios suggested by diamond anvil experiments[26,27,41]. In the first scenario, Ca-perovskite melts out first, followed by bridgmanite; ferropericlase is the liquidus phase (uppermost lower mantle to depths represented by pressures of 34–36 GPa; Extended Data Figs 3, 4). In the second scenario, Ca-perovskite melts out first, followed by ferropericlase; bridgmanite is the liquidus phase (mid-lower mantle to 60–70 GPa; Extended Data Figs 5, 6). In the third scenario, ferropericlase melts out first, followed by Ca-perovskite; bridgmanite remains the liquidus phase (lowermost mantle; Extended Data Figs 7, 8). Within each endmember scenario we run two models, representing relatively early (60–80% melting; Extended Data Figs 3a, 5a and 7a) and relatively late (80–95% melting, Extended Data Figs 3b, 5b and 7b) consumption points for the non-liquidus phases. The model tests combinations of differentiation age, differentiation degree, mixing age and mixing degree to best match the μ^{142}Nd and ε^{143}Nd data measured here and the average ε^{176}Hf measured in ref. 8. The model demands that the depleted reservoir produced in the model possesses μ^{142}Nd at least as radiogenic as the most radiogenic measured Réunion sample, and that the complimentary enriched reservoir possesses μ^{142}Nd at least as negative as the most negative measured Réunion sample. The results reported in Extended Data Figs 3e, f, 5e, f and 7e, f show coupled differentiation ages, differentiation degrees and mixing ages that best fit Nd–Hf isotopic data for a given bulk-Earth μ^{142}Nd.

All deep-mantle model setups return similar results that are primarily controlled by the point at which Ca-perovskite and ferropericlase melt out. This is because these phases fractionate Sm/Nd less than does Mg-perovskite, so that, first, they effectively buffer large changes in melt and residual Sm/Nd ratio, and second, the points at which they are consumed are marked by large changes in bulk distribution coefficients. However, regardless of this effect, best-fit differentiation models tend to be mid-Hadean in age (for example, 4.4 Gyr to 4.3 Gyr for bulk Earth μ^{142}Nd = 0 and chondritic Sm/Nd) and mixing tends to be very recent and involve only small proportions of melt (enriched reservoir, for example, 10–25% for bulk Earth μ^{142}Nd = 0 and chondritic Sm/Nd). Earlier differentiation ages in best-fit models may be correlated with shallower depths of melting (Extended Data Figs 3, 4), whereas melting in the deepest mantle tends to return somewhat younger ages (Extended Data Figs 7, 8). In general, best-fit models return early Hadean differentiation ages for modestly negative (to −10) or positive (to +5) bulk Earth μ^{142}Nd; however, the misfit to Réunion data is at its least when such early differentiation happens in the deepest mantle (Extended Data Figs 7, 8). However, internal variables (differentiation age, differentiation and mixing degrees) of the best-fit scenarios are mutually compensatory (for example, forcing a change in differentiation degree is compensated by changes in the ages of differentiation and mixing and the degree of mixing) and therefore do not offer unique solutions, even within the acceptable misfit threshold discussed above (around 21%). This condition also leads to coarseness in the model curves because the model is forced to choose between individual combinations of variables that have relatively coarse resolution (for example, 5-Myr spacing of differentiation age) lying at our maximum computational capacity. However, the general ages and degrees mentioned above represent the least-misfit scenarios and generally exclude differentiation ages in the late Hadean.

The degree of melting returned by the model closely follows phase-consumption points, as discussed above, and these can be artificially placed anywhere in the model. However, it seems unlikely that they should exist anywhere with less than 60% melting, as all three mineral phases are present in diamond-anvil melting experiments with up to 77% melt and as little as 47% melt[26]. Melting in the deepest mantle (Extended Data Figs 7, 8) might require the highest degrees of melting, but it also produces the shallowest correlations between μ^{142}Nd and ε^{143}Nd, which means that geochemical hallmarks of Hadean melting under these conditions may be most easily overprinted. Correlations between μ^{142}Nd and ε^{143}Nd appear to be shallow at the scales presented in Extended Data Figs 3–8; however, they represent mixing along modern compositions for the endmembers listed in each panel, and are identical regardless of mixing age. We therefore broadly define the most likely differentiation scenarios as involving 60–90% partial melt, and discuss the implications of such a melting event in the main text.

Shallow-mantle model calculations are based on modal, accumulated fractional melting of an olivine–clinopyroxene–orthopyroxene–garnet primitive mantle with solid modal abundances of 60%–15%–15%–10% respectively, or with 25%–25%–25%–25% melting modes (the latter case being a liquid representing an equal proportion of each mineral). Partition coefficients from the experiments of ref. 30 that did not stabilize spinel are used to construct a partial melting model to determine the Sm/Nd composition of enriched and depleted domains, assuming a bulk silicate Earth Sm/Nd of 0.196 and μ^{142}Nd of −10. A best-fit melt fraction of 1.4% and differentiation age of 4.26 Gyr is calculated for a residual, depleted domain that evolves to possess a μ^{142}Nd of 0. For differentiation at the same time, but higher melt fractions (up to around 3%), enriched and depleted domains with compositions approaching μ^{142}Nd extrema of Hadean and Archean rocks[30,40] can also be matched. Similar to what is predicted by the deep-mantle models, post-Hadean mixing between these enriched and depleted domains may reproduce the μ^{142}Nd compositions of Réunion basalts. There is no *a priori* reason to prefer the upper- or lower-mantle model, and each has their own distinct implications; however, an upper-mantle model similar to the one we present is supported by the more detailed modelling in ref. 42.

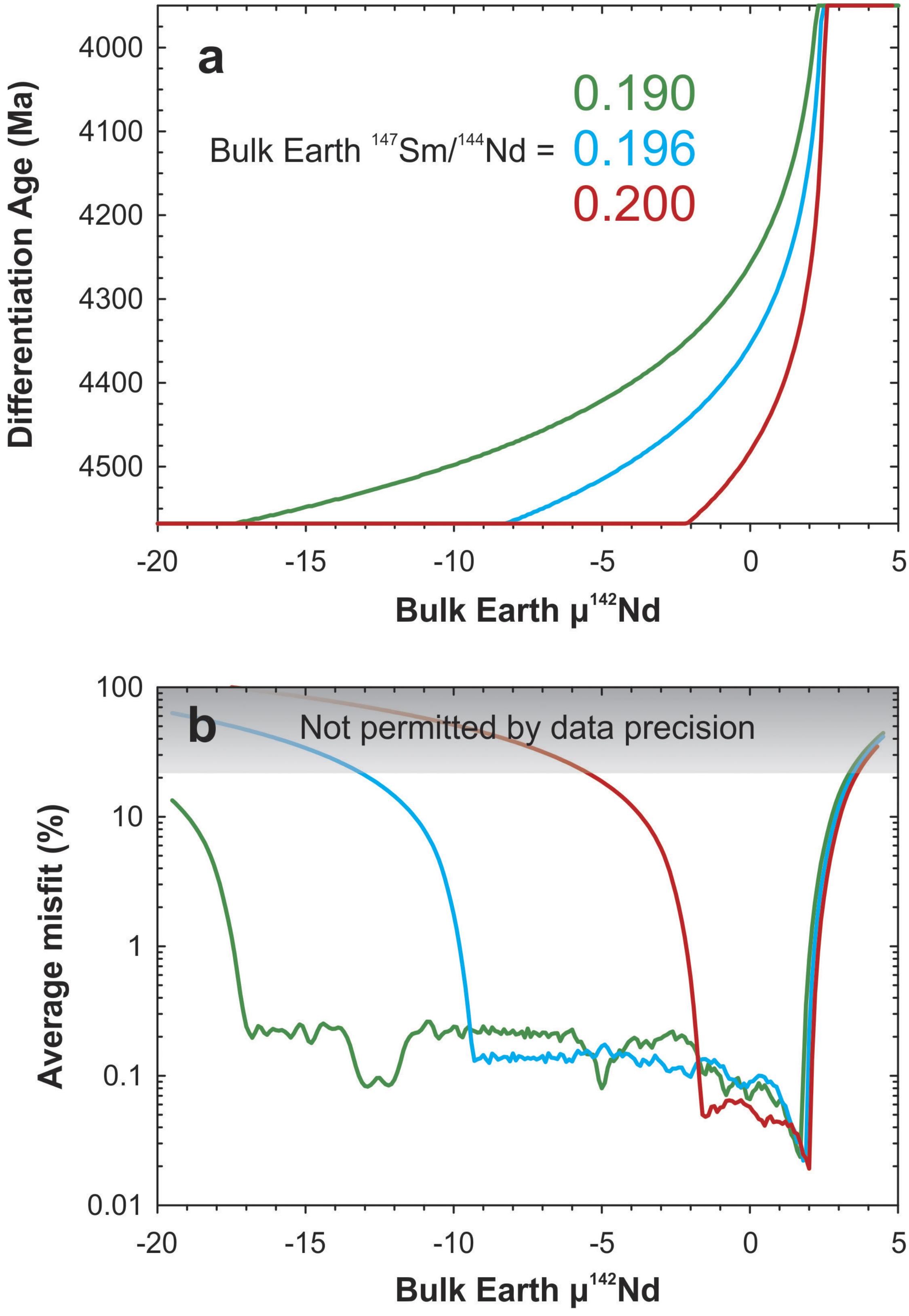
Differentiation of the Réunion source. In the main text, we discuss a primary scenario in which large-scale silicate melting in the deep mantle produces complementary trace-element-enriched and -depleted reservoirs that give rise to Réunion's μ^{142}Nd signature. The modelling we present permits several variations on this hypothesized process. First, given that our consideration of partial melting is quantitatively inverse of an analogous consideration of fractional crystallization, our numerical model is compatible with a scenario in which a melt with bulk silicate Earth composition (μ^{142}Nd as specified in the model) solidifies to a $1 - F$ (where F is the fraction of melt) degree. In other words, the model is broadly compatible with 10–40% fractional crystallization of such a melt; there is no way to distinguish between these scenarios in the numerical consideration alone. Because early fractional solids derived from such a melt may have highly distinct modal mineralogies[26], one way in which to differentiate between these two scenarios through geochemistry could be to use trace-element ratios and corresponding isotope systems that are most highly fractionated by early-crystallized phases. The petrological and geophysical implications of such a scenario have been discussed in some detail[43].

Given the possibility that one or both of the isotope domains discussed in the main text was physically solid, the trace-element-enriched endmember (that is, negative μ^{142}Nd) might represent recycled crustal material that was returned to the Réunion source. Given current knowledge of the isotopic composition of oceanic and continental crust, this notion would seem incompatible with the homogenous, relatively unradiogenic Sr–^{143}Nd–Pb–Os isotopic composition of Réunion OIB. However, given crust with lower Re/Os and (U+Th)/He ratios than modern crust, a mixing scenario similar to that outlined in the text could explain the isotopic variability of samples examined here. Rhenium is somewhat more volatile and fluid-mobile than Os, meaning that quantitative Re loss could occur upon subaerial eruption or upon subduction. Alternatively, pervasively higher mantle temperatures in the Hadean might mean that highly siderophile elements behave differently during crust formation than would be expected in modern settings, leading to lower long-term Re/Os ratios in ancient crust and less radiogenic ^{187}Os/^{188}Os ratios in recycled ancient crust. Efficient, post-Hadean mixing would decouple μ^{142}Nd from the long-term ingrowth of ^{187}Os in the Réunion source, consistent with the scenario outlined in the text and the data distribution observed in Fig. 2b.

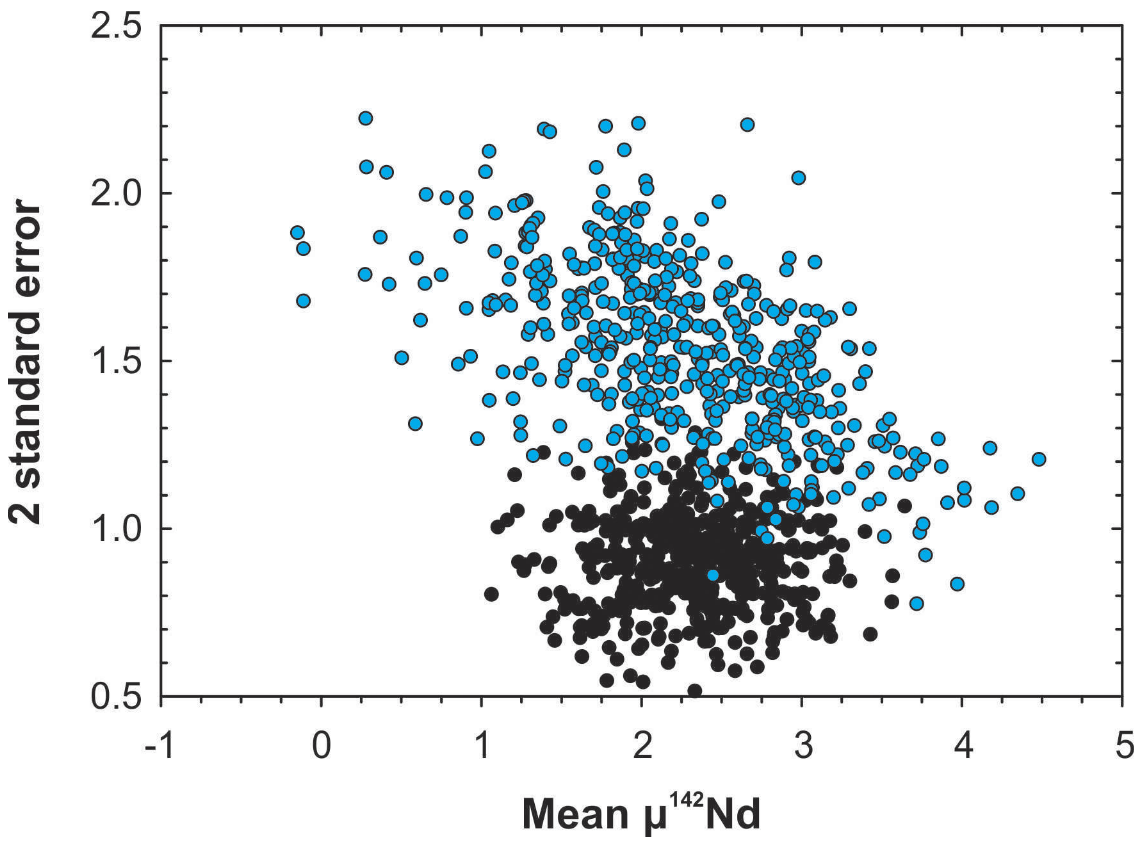
Other long-lived radiogenic isotope systems, which may be similarly decoupled from ^{142}Nd/^{144}Nd, could have behaved similarly: efficient mixing of isotope compositions and parent–daughter isotope ratios after ^{146}Sm was extinct would give rise to a homogenous isotopic composition that is decoupled from μ^{142}Nd. By contrast, given that μ^{142}Nd correlates with ^{3}He/^{4}He (Fig. 2a), recycling of Hadean crust that had been efficiently degassed may produce a situation in which the ^{3}He/^{4}He signature of the enriched crustal material is overprinted by the helium-isotope signature of the depleted mantle residuum. On the other hand, the rate of ^{4}He ingrowth would be determined primarily by the addition of uranium and thorium from the recycled crust, because it would have very high concentrations of these elements relative to the depleted residuum. This presents a situation in which μ^{142}Nd and ^{3}He/^{4}He would initially be uncorrelated, but would evolve to be correlated with a slope moderated by $(U+Th)_{crust}/He_{residuum}$. In addition, if the depleted residuum would have otherwise evolved to possess extremely high ^{3}He/^{4}He, as is observed in some OIBs, then addition of uranium and thorium from recycled crust might have produced the more moderate ^{3}He/^{4}He signature that is characteristic of Réunion OIBs. Given the likelihood that Archaean crust, like modern crust, is relatively enriched in U and Th, the concentration of He in the residual mantle source to hotspots may be the most likely governor of this relationship.

Data availability. All of the data generated here are available in the Supplementary Tables and Source Data files. The Nd isotope data are also displayed in Figs 1–4 and Extended Data Figs 4, 6, 8 and 10; previously published He and Os data are shown in Fig. 2. Extended Data Figs 1, 2 and 9 contain only the outcomes of numerical models, with no original data. All data are also available for download from the EarthChem database (www.earthchem.org, doi:10.1594/IEDA/100740).

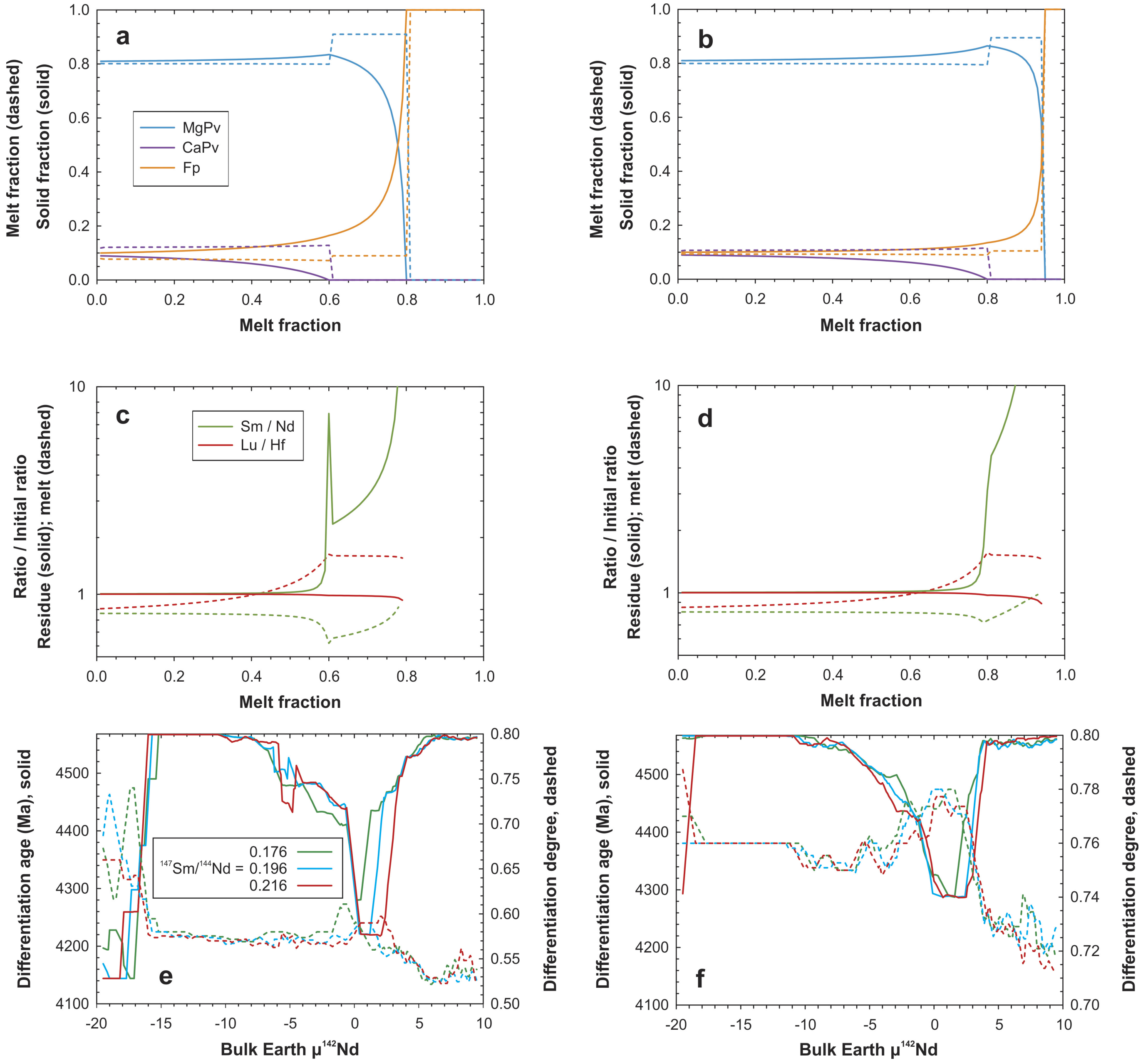
34. de Leeuw, G. A. M., Ellam, R. M., Stuart, F. M. & Carlson, R. W. ^{142}Nd/^{144}Nd inferences on the nature and origin of the source of high ^{3}He/^{4}He magmas. *Earth Planet. Sci. Lett.* **472,** 62–68 (2017).
35. Garçon, M., Boyet, M., Carlson, R. W., Horan, M. F., Auclair, D. & Mock, T. D. Factors influencing the precision and accuracy of Nd isotope measurements by thermal ionization mass spectrometery. *Chem. Geol.* **476,** 493–514 (2018).
36. Peters, B. J., Carlson, R. W., Day, J. M. D. & Horan, M. F. Nd isotope data for Réunion Island basalts and cumulate xenoliths. *EarthChem* http://dx.doi.org/10.1594/IEDA/100740 (2018).
37. Burkhardt, C. *et al.* A nucleosynthetic origin for the Earth's anomalous ^{142}Nd composition. *Nature* **537,** 394–398 (2016).
38. Render, J., Fischer-Gödde, M., Burkhardt, C. & Kleine, T. The cosmic molybdenum-neodymium isotope correlation and the building material of the Earth. *Geochem. Persp. Lett.* **3,** 170–178 (2017).
39. Hilton, D. R. *et al.* Helium isotopes at Rungwe volcanic province, Tanzania, and the origin of East African plateau. *Geophys. Res. Lett.* **38,** (2011).
40. O'Neil, J., Rizo, H., Boyet, M., Carlson, R. W. & Rosing, M. T. Geochemistry and Nd isotopic characteristics of Earth's Hadean mantle and primitive crust. *Earth Planet. Sci. Lett.* **442,** 194–205 (2016).
41. Nomura, R. *et al.* Spin crossover and iron-rich silicate melt in the Earth's deep mantle. *Nature* **473,** 199–202 (2011).
42. Brown, S. M., Elkins-Tanton, L. T. & Walker, R. J. Effects of magma ocean crystallization and overturn on the development of ^{142}Nd and ^{182}W isotopic heterogeneities in the primordial mantle. *Earth Planet. Sci. Lett.* **408,** 319–330 (2014).
43. Labrosse, S., Hernlund, J. W. & Hirose, K. in *The Early Earth: Accretion and Differentiation* (eds Badro, J. & Walter, M.) 123–142 (AGU, 2015).
44. Cipriani, A., Bonatti, E. & Carlson, R. W. Nonchondritic ^{142}Nd in suboceanic mantle peridotites. *Geochem. Geophys. Geosyst.* **12,** Q03006 (2011).



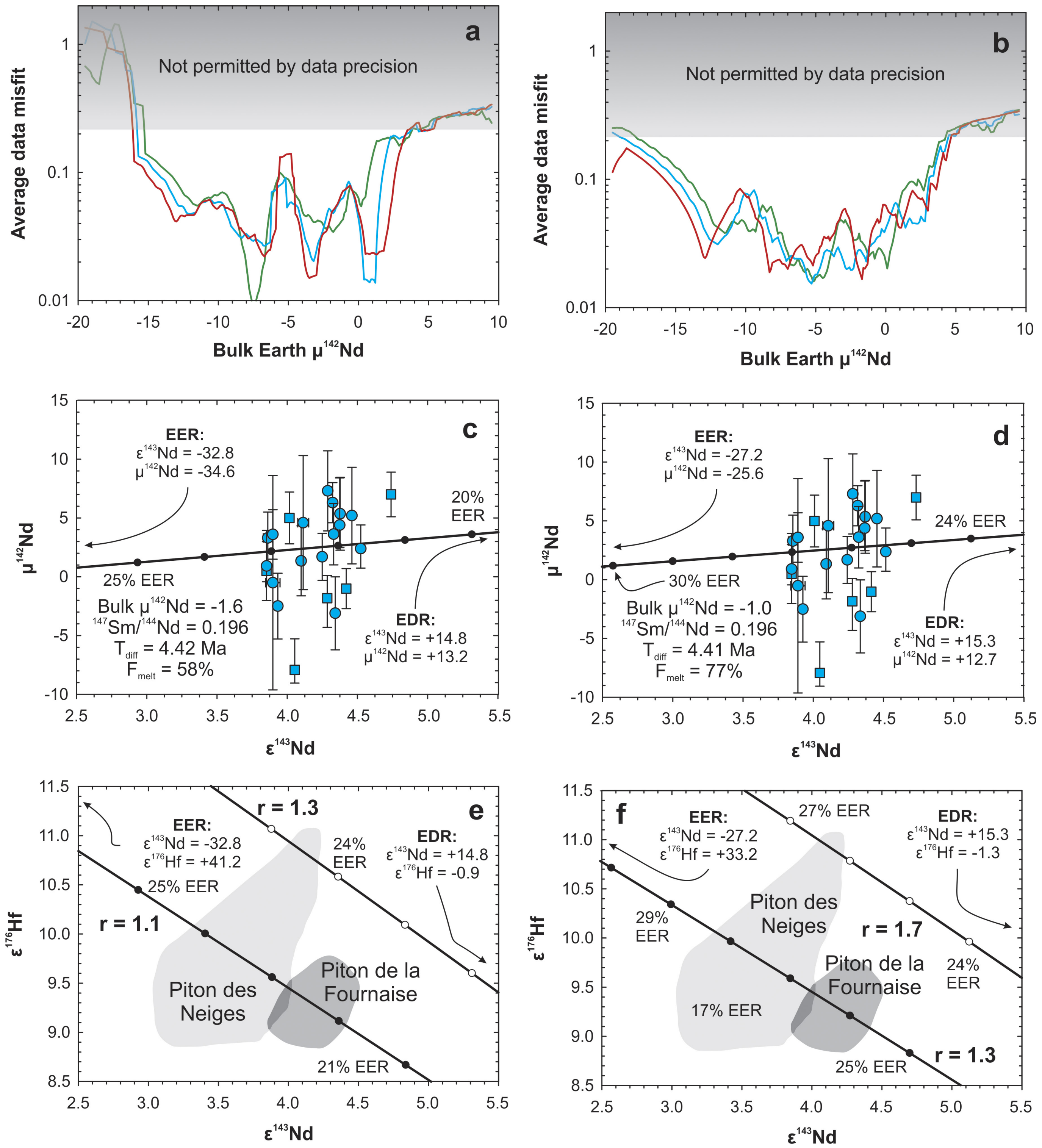
Extended Data Figure 1 | Details of possible single-stage differentiation models. For example, the model shown in Fig. 3. **a**, Required differentiation ages (*y* axis) for given bulk-Earth μ^{142}Nd values (*x* axis) and ^{147}Sm/^{144}Nd ratios (coloured lines) to best match Réunion average μ^{142}Nd and ε^{143}Nd. Given external precisions as a proportion of data range, minimum bulk-Earth μ^{142}Nd values permitted by the Réunion data are indicated for various bulk Earth ^{147}Sm/^{144}Nd ratios (see text for details). **b**, Misfit of each scenario to Réunion data. Misfits are calculated as a moving average with a 1 p.p.m. μ^{142}Nd window; coarseness among low misfits is a consequence of model resolution. Bulk Earth μ^{142}Nd values are preserved vertically across the panels. For example, assuming a bulk Earth with Sm/Nd = 0.19 and μ^{142}Nd = −5, the differentiation age that produces the best fit to Réunion μ^{142}Nd and ε^{143}Nd is around 4,430 million years, and the total misfit to Réunion data is less than 1%, meaning that this scenario is permitted within the precision of the data. Ma, million years.



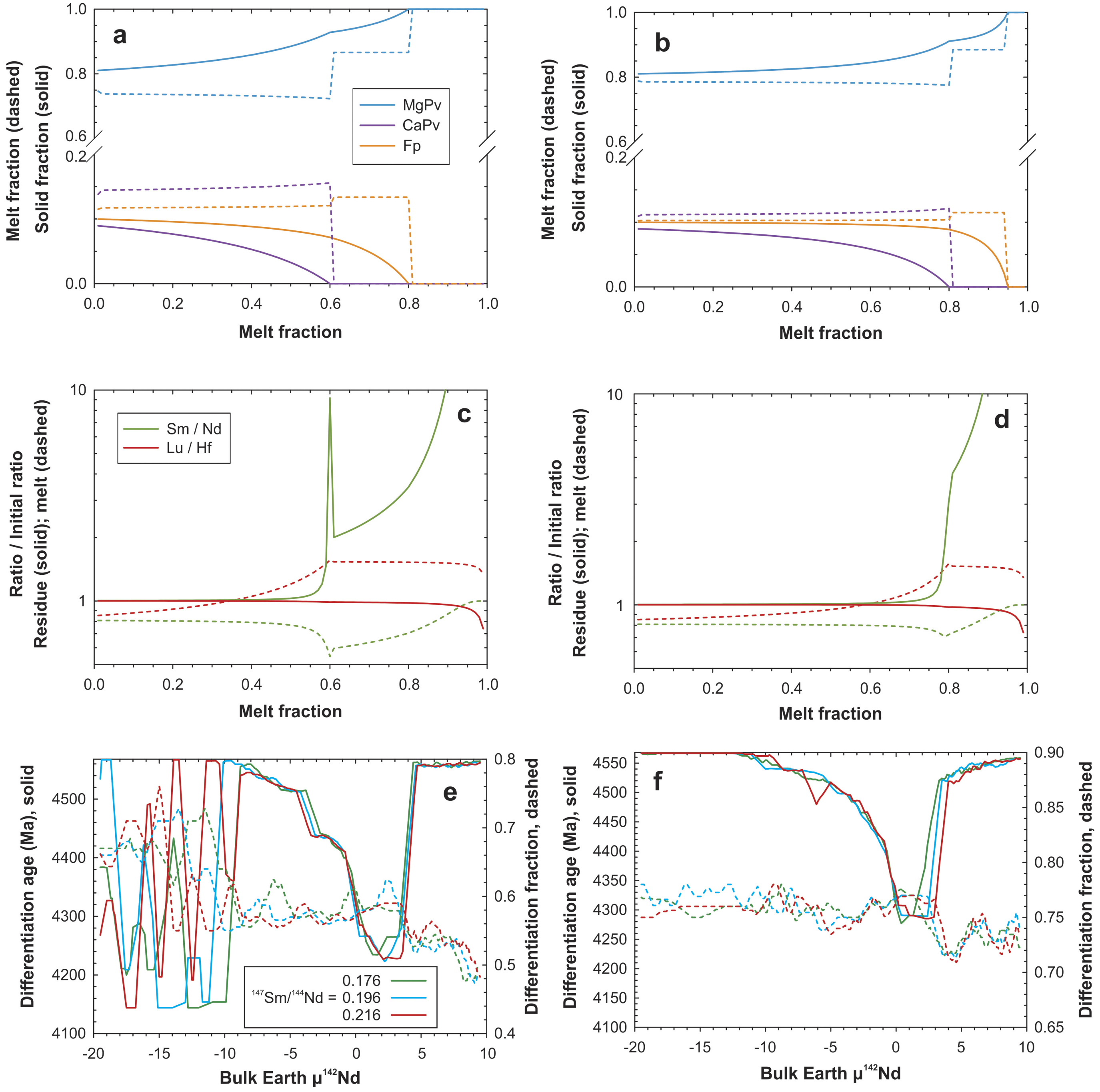
Extended Data Figure 2 | Hypothesis testing, showing that the spread of data among Réunion igneous rocks is greater than would be expected from analytical precision alone. Black dots show arithmetic means and s.e.m. for randomly generated μ^{142}Nd values, with a sample mean equal to the Réunion μ^{142}Nd mean and standard deviation equal to external precision. Blue dots show arithmetic means and s.e.m. for randomly selected Réunion sample data with replacement after selection (that is, a given sample can be chosen one or more times). Each point (black or blue) represents a sample size of $n = 22$, identical to the number of samples analysed in this study. The two simulated datasets (black and blue dots) do not strongly overlap, implying that the total variation in Réunion μ^{142}Nd is greater than would be expected from external precision alone. See text for further details and interpretation; an example simulation is given in the Source Data.



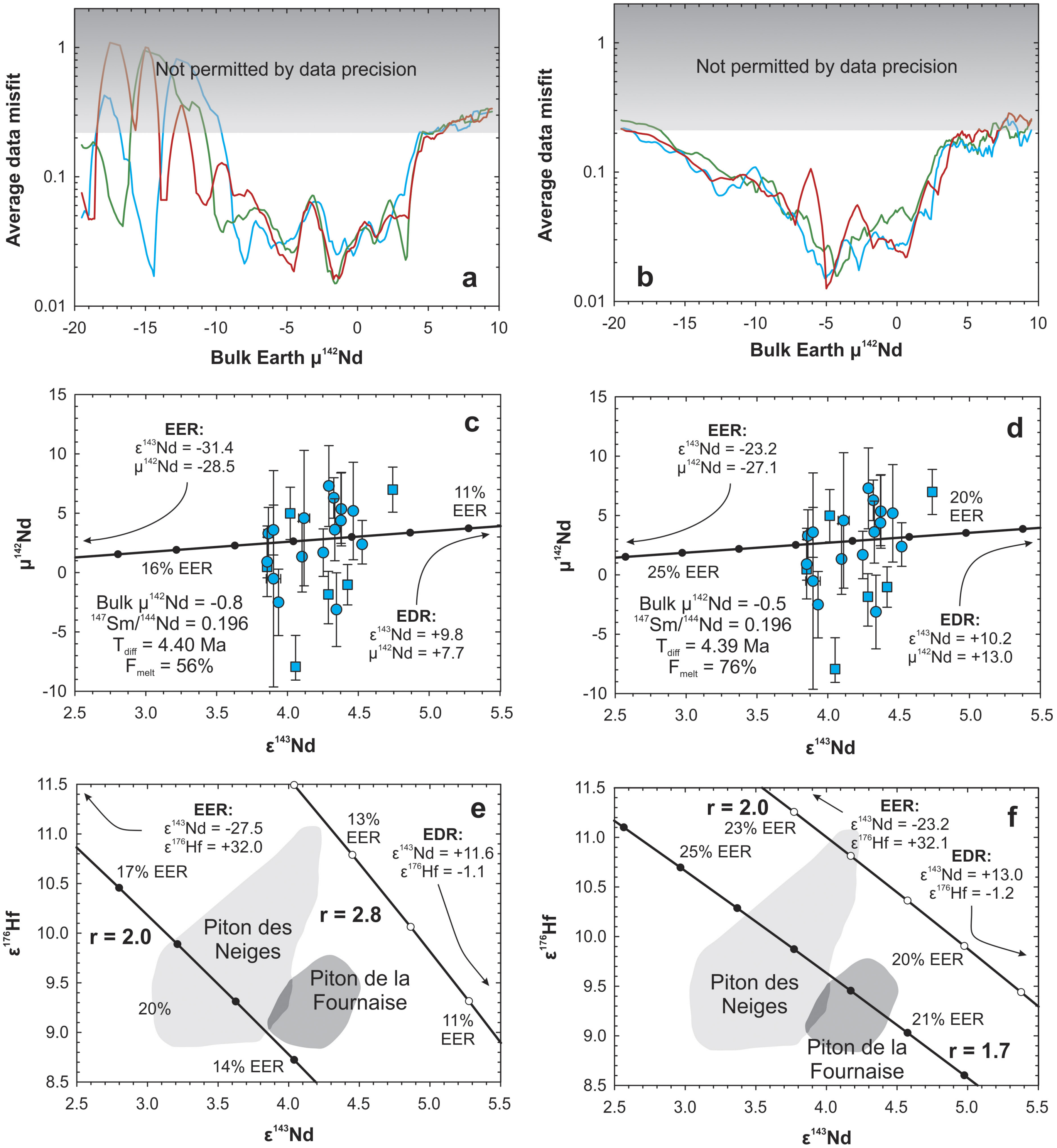
Extended Data Figure 3 | Isotope evolution model for differentiation occurring under pressures representative of the modern uppermost lower mantle. At this depth, pressures are less than 34–36 GPa (ref. 27); ferropericlase (Fp) is the liquidus phase and Ca-perovskite (Ca-Pv) melts out before bridgmanite (Mg-perovskite, Mg-Pv). Results from a single model run are shown vertically in each column, with melt-out points set to 60% (Ca-Pv) and 80% (Mg-Pv) of melting (left panels), and to 80% (Ca-Pv) and 95% (Mg-Pv) of melting (right panels). **a**, **b**, Mantle and melt modes. **c**, **d**, Changes in Sm/Nd and Lu/Hf through accumulated fractional melting. **e**, **f**, Best-fit differentiation ages and melting degrees for given bulk-Earth μ^{142}Nd and Sm/Nd values, given as moving averages around a 1 p.p.m. window. Bridgmanite and Ca-perovskite partition coefficients are from ref. 21; partitioning of Hf and rare earth elements in ferropericlase is assumed to be nil.



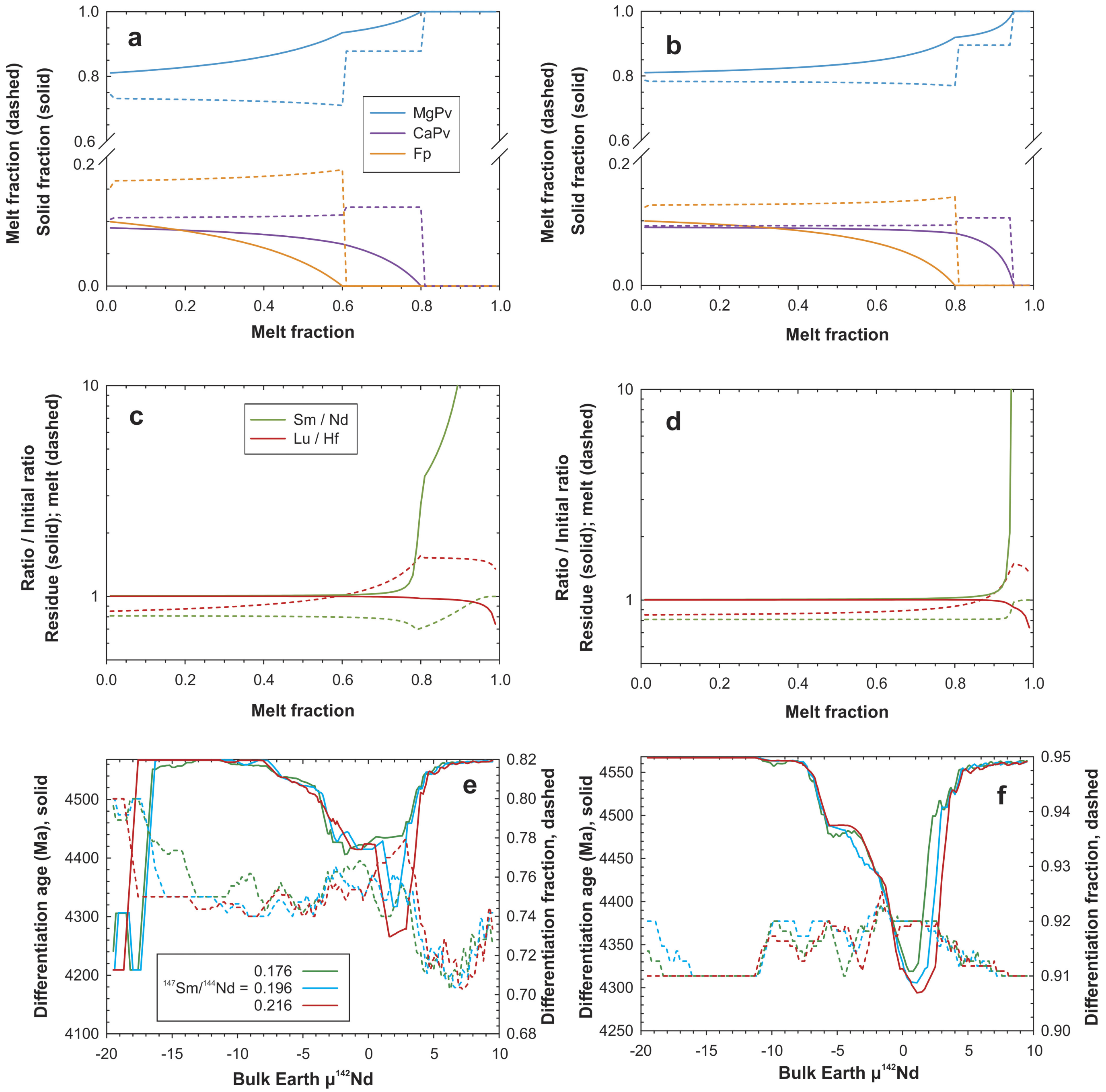
Extended Data Figure 4 | Isotope evolution model for differentiation occurring under pressures representative of the modern uppermost lower mantle. Continuing on from Extended Data Fig. 3, results from a single model run are shown vertically in each column, with melt-out points set to 60% and 80% of melting (left panels), and to 80% and 95% of melting (right panels). **a**, **b**, Misfit to data averages for scenarios illustrated in Extended Data Fig. 3e, f. **c**, **d**, Selected model fit to Réunion μ^{142}Nd–ε^{143}Nd data. **e**, **f**, Same model as in panels **c**, **d**, with fit to Réunion ε^{143}Nd–ε^{176}Hf (ref. 8) data. Best-fit models return slightly earlier differentiation ages compared with deeper-mantle models (Extended Data Figs 3, 5 and 7) and require a lower degree of differentiation than does melting occurring at lowest mantle pressures (Extended Data Figs 7, 8). Differentiation at these relatively low pressures produces the steepest negative correlations between ε^{176}Hf and ε^{143}Nd compared with differentiation at higher pressures. Correlations between μ^{142}Nd and ε^{143}Nd appear to be shallow at the scales presented in panels **c** and **d**; however, they reflect the small *x*-axis scale that best represents the distribution of Réunion data. Réunion Nd-isotope data in panels **c** and **d** can be found in Supplementary Table 1 or in Source Data.



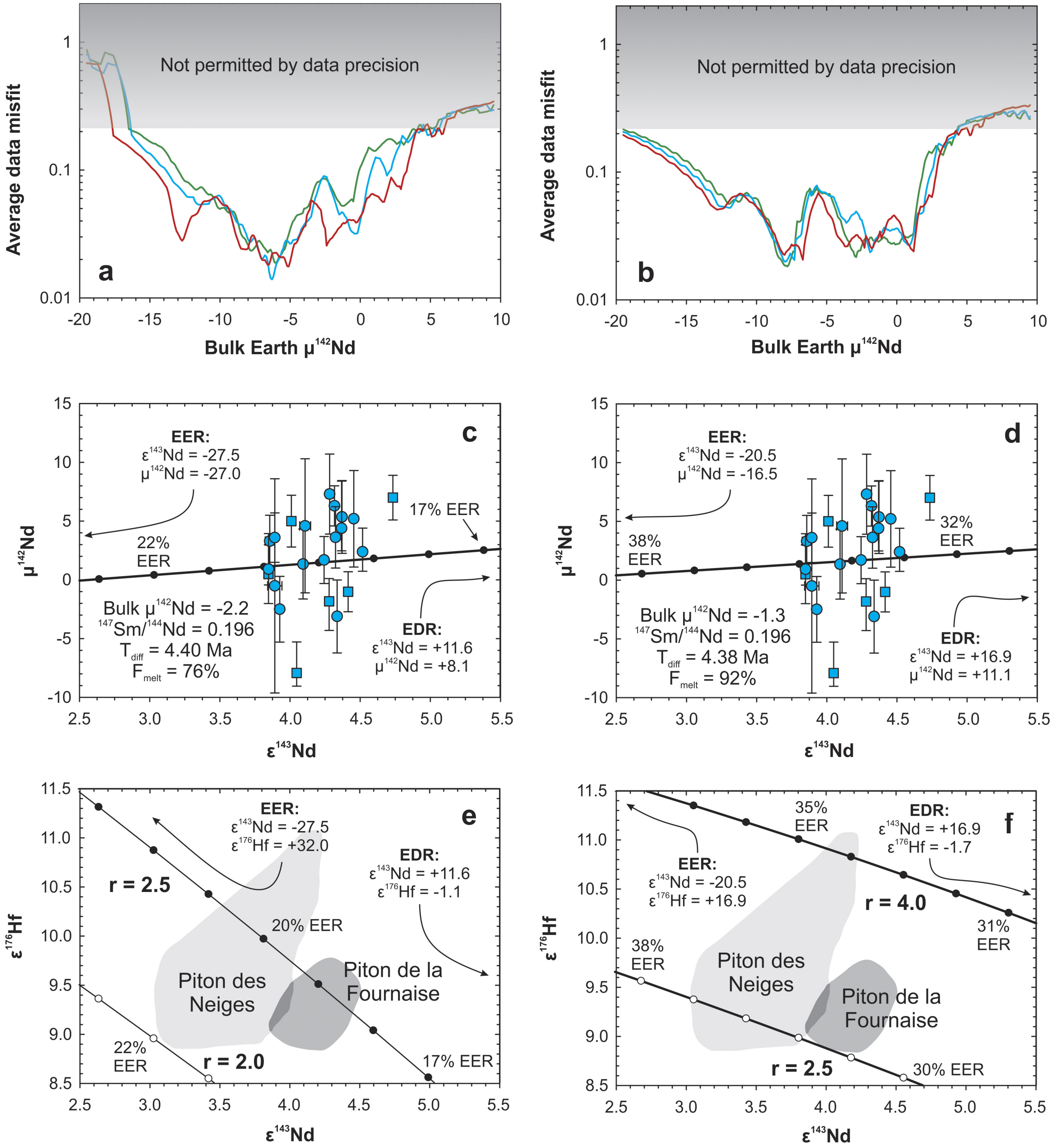
Extended Data Figure 5 | Isotope-evolution model for differentiation occurring under pressures representative of the modern mid-lower mantle. At these pressures (36–60 GPa; ref. 26), bridgmanite is the liquidus phase and Ca-perovskite melts out before ferropericlase. Left panels: 60% (Ca-Pv) and 85% (Fp) melt-out points; right panels: 80% (Ca-Pv) and 95% (Fp) melt-out points. **a**, **b**, Mantle and melt modes. **c**, **d**, Changes in Sm/Nd and Lu/Hf through accumulated fractional melting. **e**, **f**, Best-fit differentiation ages and melting degrees for given bulk-Earth μ^{142}Nd and Sm/Nd values, given as moving averages around a 1 p.p.m. window. A relatively constant melting degree (71–77%) is predicted over a range of bulk Earth μ^{142}Nd values and differentiation ages when Ca-perovskite and ferropericlase are consumed late during melting (panel **f**), making these model conditions the most adaptable to possible conditions of Hadean mantle domain formation.



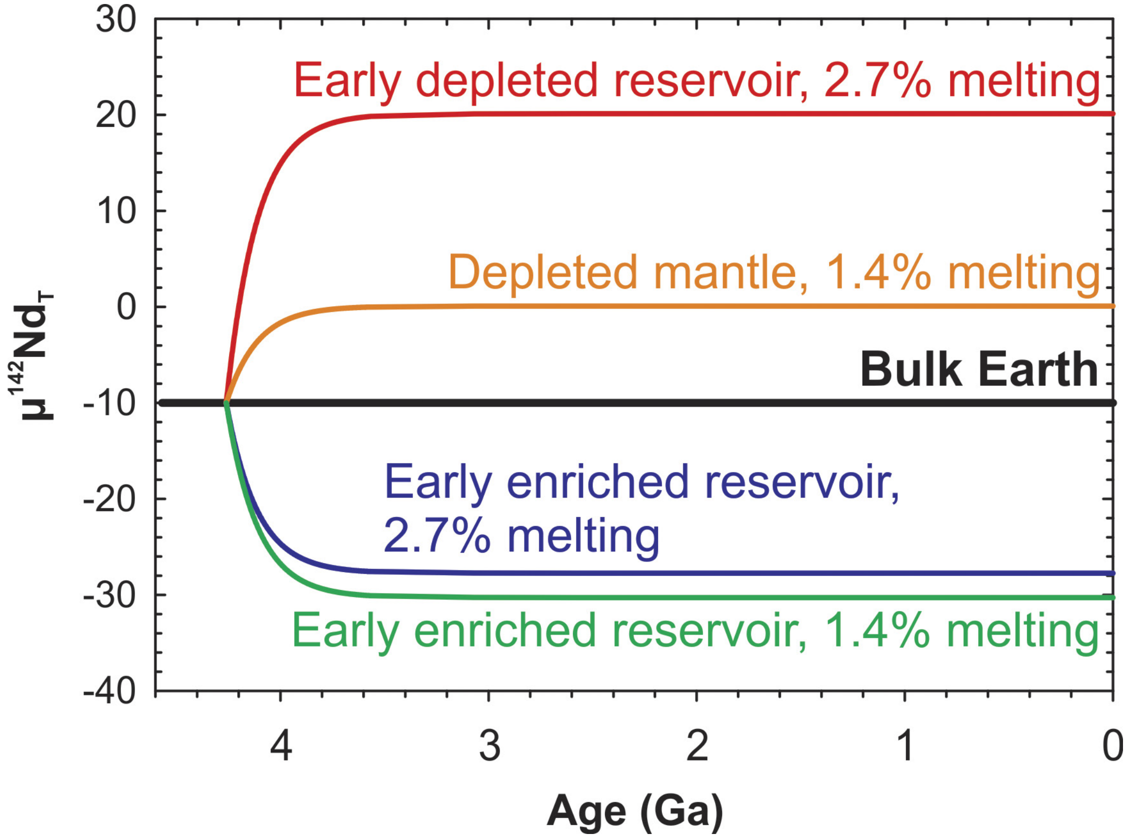
Extended Data Figure 6 | Isotope-evolution model for differentiation occurring under pressures representative of the modern mid-lower mantle. Continuing on from Extended Data Fig. 5, left panels: 60% and 85% melt-out points; right panels: 80% and 95% melt-out points. **a**, **b**, Misfit to data averages for scenarios illustrated in Extended Data Fig. 5e, f. **c**, **d**, Selected model fit to Réunion μ^{142}Nd–ε^{143}Nd data. **e**, **f**, Same model as in panels **c**, **d**, with fit to ε^{143}Nd–ε^{176}Hf (ref. 8) data. Somewhat later preferred differentiation times (around 4.40 Gyr ago) produce slightly shallower ε^{176}Hf–ε^{143}Nd correlations compared with those observed in Extended Data Figs 3 and 4, and required contributions from the EER (**c**, **d**) are less than for other pressure conditions (Extended Data Figs 3, 4 and Extended Data Figs 7, 8) Correlations between μ^{142}Nd and ε^{143}Nd appear shallow at the scales presented in panels **c** and **d**; however, they reflect the small *x*-axis scale that best represents the distribution of Réunion data. Réunion Nd isotope data in panels **c** and **d** can be found in Supplementary Table 1 and in the Source Data.



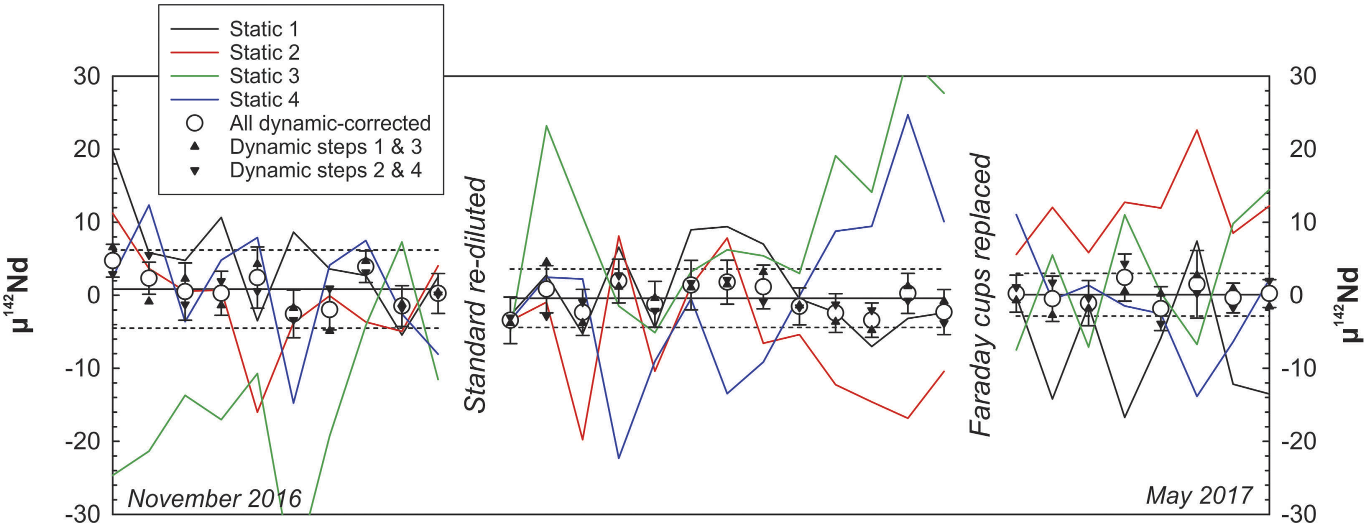
Extended Data Figure 7 | Isotope-evolution model for differentiation occurring under pressures representative of the modern lowest mantle. At these pressures (greater than 60 GPa; ref. 26), bridgmanite is the liquidus phase and ferropericlase melts out before Ca-perovskite. Left panels: 60% and 85% melt-out points; right panels: 80% and 95% melt-out points. **a**, **b**, Mantle and melt modes. **c**, **d**, Changes in Sm/Nd and Lu/Hf through accumulated fractional melting. **e**, **f**, Best-fit differentiation ages and melting degrees for given bulk-Earth μ^{142}Nd and Sm/Nd values, given as moving averages around a 1 p.p.m. window. In this highest pressure range, the model demands the highest degree of mantle melting (greater than 90% for late consumption of ferropericlase and Ca-perovskite), but returns the shallowest correlations between ε^{143}Nd, μ^{142}Nd and ε^{176}Hf and under the same conditions (see Extended Data Fig. 8).



Extended Data Figure 8 | Isotope-evolution model for differentiation occurring under pressures representative of the modern lowest mantle. Continuing on from Extended Data Fig. 7, left panels: 60% and 85% melt-out points; right panels: 80% and 95% melt-out points. **a**, **b**, Misfit to data averages for scenarios illustrated in Extended Data Fig. 7e, f. **c**, **d**, Selected model fit to Réunion μ^{142}Nd–ε^{143}Nd data. **e**, **f**, Same model as in panels **c**, **d**, with fit to ε^{143}Nd–ε^{176}Hf (ref. 8) data. Deep-mantle Hadean domains formed under these conditions are possibly the most susceptible to any overprinting that would mask correlations between μ^{142}Nd and long-lived radiogenic isotope systems. Correlations between μ^{142}Nd and ε^{143}Nd appear to be shallow at the scales presented in panels **c** and **d**; however, they reflect the small *x*-axis scale that best represents the distribution of Réunion data. Réunion Nd isotope data in panels **c** and **d** can be found in Supplementary Table 1 or in the Source Data.



Extended Data Figure 9 | A non-unique, upper-mantle differentiation scenario occurring in a bulk Earth with present-day values for $^{147}Sm/^{144}Nd$ of 0.196 and $\mu^{142}Nd$ of −10. Melting of 1.4% by mass at an age of 4.26 Gyr will form an early enriched reservoir (green line) that will evolve to possess a $\mu^{142}Nd$ of about −31, and will leave behind a residual depleted mantle (orange line) that will evolve to possess a $\mu^{142}Nd$ of 0; this latter value is identical to the terrestrial standard JNdi and identical to estimates for the composition of the upper mantle[44]. If melting locally progresses to 2.7%, it will produce an enriched reservoir (blue line) with a $\mu^{142}Nd$ of about −28 and a conjugate depleted reservoir (red line) with a $\mu^{142}Nd$ of about +20, which may subsequently mix to produce a $\mu^{142}Nd$ composition similar to that of the Réunion source. Full model inputs are described in the Methods. Note that, unlike in Fig. 4a, the *y*-axis is normalized to the $^{142}Nd/^{144}Nd$ composition of the bulk Earth at a given *x*-axis time ($\mu^{142}Nd_T$) as opposed to the $^{142}Nd/^{144}Nd$ composition of the modern bulk Earth.



Extended Data Figure 10 | Run information for the terrestrial neodymium standard JNdi throughout the analysis campaign, normalized to average ^{142}Nd/^{144}Nd before correction. Analyses typically comprised 600–900 cycles of four steps each; steps 1 and 3, and steps 2 and 4, were each paired to calculate two dynamic-corrected ^{142}Nd/^{144}Nd ratios per cycle. Dynamic-corrected values correct for uncertainties in static measurements that are introduced by variable Faraday cup efficiencies. See Methods for details. Dynamic-corrected symbols represent averages and 2 s.e.m. of all dynamic-corrected ratios from that run.

doi:10.1038/nature25783

Evolutionary history resolves global organization of root functional traits

Zeqing Ma[1]*, Dali Guo[1]‡*, Xingliang Xu[1], Mingzhen Lu[2], Richard D. Bardgett[3], David M. Eissenstat[4], M. Luke McCormack[1,5] & Lars O. Hedin[2]*

Plant roots have greatly diversified in form and function since the emergence of the first land plants[1,2], but the global organization of functional traits in roots remains poorly understood[3,4]. Here we analyse a global dataset of 10 functionally important root traits in metabolically active first-order roots, collected from 369 species distributed across the natural plant communities of 7 biomes. Our results identify a high degree of organization of root traits across species and biomes, and reveal a pattern that differs from expectations based on previous studies[5,6] of leaf traits. Root diameter exerts the strongest influence on root trait variation across plant species, growth forms and biomes. Our analysis suggests that plants have evolved thinner roots since they first emerged in land ecosystems, which has enabled them to markedly improve their efficiency of soil exploration per unit of carbon invested and to reduce their dependence on symbiotic mycorrhizal fungi. We also found that diversity in root morphological traits is greatest in the tropics, where plant diversity is highest and many ancestral phylogenetic groups are preserved. Diversity in root morphology declines sharply across the sequence of tropical, temperate and desert biomes, presumably owing to changes in resource supply caused by seasonally inhospitable abiotic conditions. Our results suggest that root traits have evolved along a spectrum bounded by two contrasting strategies of root life: an ancestral 'conservative' strategy in which plants with thick roots depend on symbiosis with mycorrhizal fungi for soil resources and a more-derived 'opportunistic' strategy in which thin roots enable plants to more efficiently leverage photosynthetic carbon for soil exploration. These findings imply that innovations of belowground traits have had an important role in preparing plants to colonize new habitats, and in generating biodiversity within and across biomes.

Recent efforts to understand how functional traits are organized across land plants have revealed notable patterns across the leaf economic spectrum[5,6], but whether such a high degree of organization is also seen in root traits remains controversial[4,7]. A key factor that has limited progress has been the paucity of data on root traits across plant species and biomes, as roots are difficult to sample and characterize[8,9]. Yet, roots are vital for the ability of plants to acquire nutrients and water—two functions of fundamental importance to whole-plant performance and for predicting how plants respond to elevated CO_2 levels and to climate change[10–12].

Roots face ecological and physiological challenges that differ fundamentally from those encountered by leaves. Roots must compete for and acquire nutrients and water in environments that greatly vary across global biomes, with biophysical conditions ranging from relatively stable (for example, tropical rainforests) to highly seasonal (for example, deserts or boreal forests). The high diversity that exists in root form and function, and in the degree of association with symbiotic mycorrhizal fungi, raises a fundamental question: how are root traits organized across the diverse taxa that inhabit different ecological conditions worldwide?

Here we propose a model of root trait organization that is functionally decoupled from the leaf economic spectrum and that derives from the phylogenetic history of root diameter and its evolutionary consequences for plant resource acquisition.

We evaluated a species- and biome-specific dataset of 10 root traits in 3 major categories[3,13] (morphology, physiology and mycorrhizal association; Supplementary Information, note 1), collected from over 1,200 individual plants of 369 species (from 210 genera and 79 families), distributed across 7 major biomes and 3 continents (Extended Data Table 1). The observations in our dataset: (i) derive solely from native plant communities with natural soil and nutrient conditions; (ii) focus on first-order roots (the most distal and absorptive roots of the branching system) that are subject to strong selection by the local environment[8,9,14]; (iii) accurately identify species and root order (that is, measure of branching hierarchy[8]) in mixed-species ecosystems, by tracing roots to parent trees[15]; and (iv) apply consistent analytical methods to trait measures across all species and biomes. We collected 94% of the total observations used in our dataset (see Methods).

We first investigated whether first-order root traits are globally organized in a manner analogous to the leaf economic spectrum[5,6], a composite axis of trait variation that ranges from nitrogen-rich leaves with high specific leaf area and short leaf lifespan to nitrogen-poor leaves with low specific leaf area and long leaf lifespan. In roots, nitrogen supports metabolic activity, including nutrient and water transport, enzyme functioning and mycorrhizal symbiosis[16]. As a result, nitrogen has previously been proposed to serve a similarly central role in the trait organization of roots, with high levels of nitrogen in roots occurring in species with high levels of nitrogen in their leaves, rapid growth and short root lifespans[4,17].

Our results do not support the idea of an analogous organizing role for nitrogen in a global root economic spectrum, expanding on similar previous conclusions drawn from taxonomically and geographically smaller datasets[4,18,19]. First, a principal component analysis failed to identify root nitrogen, which is analogous to leaf nitrogen, as a significant contributor to the primary axis of trait variation (Extended Data Fig. 1 and Extended Data Table 2). Instead, root traits were most strongly explained by root diameter and by a group of traits associated with root construction and mycorrhizal association (principal component 1, 46%; Extended Data Fig. 1).

Second, root nitrogen was not correlated to specific root length (SRL, the length of root per unit of biomass invested) in a manner analogous to the relationship between specific leaf area and leaf nitrogen

[1]Center for Forest Ecosystem Studies and Qianyanzhou Ecological Station, Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China. [2]Department of Ecology and Evolutionary Biology, Princeton University, New Jersey 08544, USA. [3]School of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PT, UK. [4]Department of Ecosystem Science and Management, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [5]Department of Plant and Microbial Biology, University of Minnesota, St Paul, Minnesota 55108, USA.

*These authors contributed equally to this work.

‡Deceased.

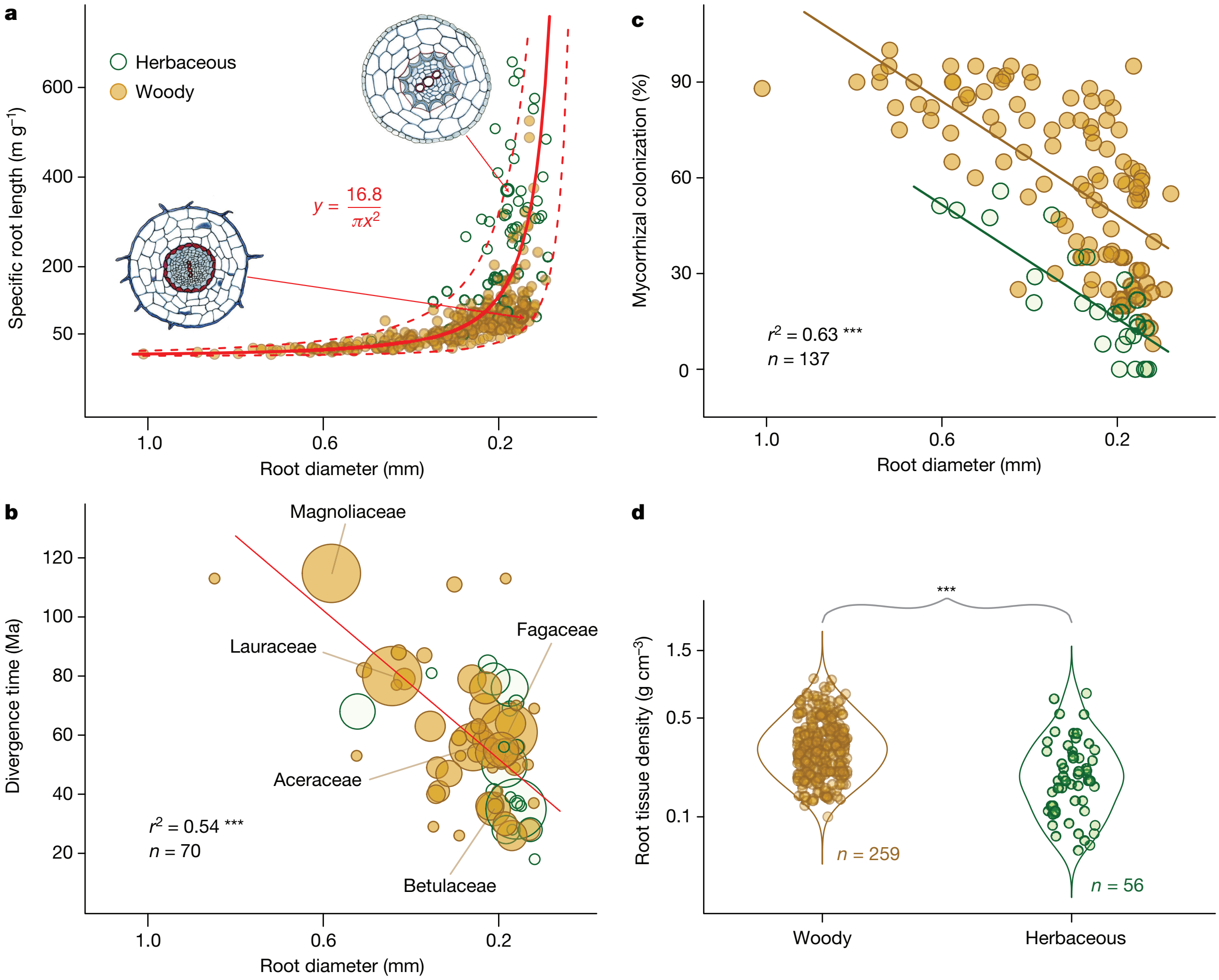


Figure 1 | Root trait dimensions organized by root diameter and growth form. a–c, At the species level, the diameter of first-order roots is inversely correlated with SRL (**a**, $n = 323$ species), positively related to the evolutionary time of divergence of major taxonomic groups (**b**; Ma, million years ago) and positively related to the length of root (in per cent) colonized by mycorrhizal fungi (**c**). ***$P < 0.001$, linear regression. **d**, Root tissue density differs across plant growth form, with herbaceous plants (green) displaying more constrained variation than woody plants (brown) (F-test, ***$P < 0.001$; note logarithmic scale ($\log_{10}$) on the y axis). The solid red line in **a** identifies the relationship $y = 16.8/(\pi x^2)$ in which y denotes SRL and x denotes root diameter, and root tissue density is $0.25\,\mathrm{g\,cm^{-3}}$ (Supplementary Information, note 1); upper and lower dashed red lines identify tissue densities of 0.1 and $1.0\,\mathrm{g\,cm^{-3}}$, respectively. We used a linear regression weighted by number of species in **b**, and a linear regression with woody and non-woody growth forms as categorical variables in **c**. Root cross-section images in **a** are from the low root-tissue-density grass *Agropyron cristatum* (right) and the high root-tissue-density woody shrub *Rhaphiolepis indica* (left).

(Extended Data Figs 2a, b, 3). Third, *in situ* ($n = 75$) and hydroponic-based ($n = 111$) measures showed no systematic relationship between root nitrogen uptake—which is analogous to leaf photosynthetic capacity—and root diameter, SRL or plant growth form (Extended Data Figs 2c, d, 4b, d). Taken together, these results suggest that nitrogen is less important in belowground nutrient foraging than in aboveground light and CO_2 capture (Supplementary Information, note 3). Furthermore, root lifespan, which is analogous to leaf lifespan, was correlated with root diameter and SRL but explained only 14% and 17% of the respective variance ($P < 0.01$ for both, linear model; Extended Data Fig. 4a, c).

We next analysed the organizing role of root diameter in determining trait variation across plants. We found that the length of root per biomass invested (that is, the SRL) increases nonlinearly with decreasing root diameter (D) according to the allometric relationship, $\mathrm{SRL} = 16.8/(\pi D^2)$ (Fig. 1a, solid red line; Supplementary Information, note 1). This relationship indicates that as roots get thinner, plants can explore markedly greater volumes of soil per unit of carbon they invest. We also found that woody and herbaceous plants occupy different parts of the SRL versus root diameter relationship, although there is some overlap: woody plants (Fig. 1a, brown points) tend to occupy a region in which differences in root diameter have a limited effect on SRL, whereas herbaceous plants (Fig 1a, green points) reside in a region in which even small variations in diameter cause large changes in SRL.

We further found that in thin-rooted species even modest evolutionary changes in the tissue density of first-order roots can greatly alter the soil length explored per unit of carbon invested. The dashed red lines in Fig. 1a indicate the sensitivity of the relationship between SRL and diameter to changes in root tissue density across a physiologically relevant range (0.1–$1\,\mathrm{g\,cm^{-3}}$). For example, the low root density of the grass *Agropyron cristatum* enables it to explore approximately 350 m more soil per gram of biomass than can the shrub *Rhaphiolepis indica*, despite the fact that both have root diameters of approximately 0.2 mm (Fig. 1a, red arrows and cross-section images). We infer that over evolutionary time plants have used both root diameter (Fig. 1a, x axis) and tissue density (Fig. 1a, dashed lines) to influence SRL: thin and soft first-order roots have the advantage of efficient soil exploration, but incur the cost trade-off of sacrificing water conduction, tissue permanence and the ability to penetrate the soil matrix.

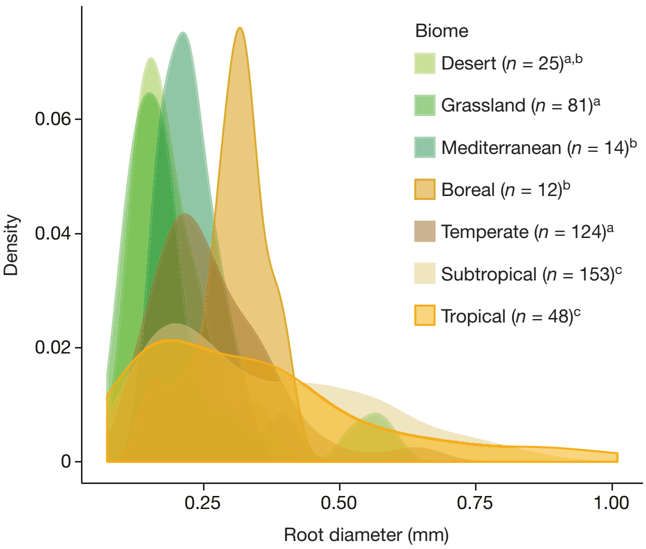


Figure 2 | Density distributions of first-order root diameter across seven biomes. The variance in root diameter declines from biomes with equable conditions (for example, tropical forests) to biomes with pronounced seasonality in soil resource supplies (for example, deserts). Numbers in brackets identify species-specific observations. Letters denote groups; pairwise differences between members of different groups are significant, Levene's variance test (for P values, see Extended Data Fig. 5c). Woody biomes are identified as shades of tan-to-yellow and non-woody biomes as shades of green.

We next investigate the role of evolutionary history in structuring the differences in root diameter across all major vascular plant families in our dataset (Fig. 1b). We found that, on average, thick roots are associated with evolutionarily ancient taxa (for example, Magnoliaceae) and that thin roots are increasingly common in taxa that have recently diverged from their ancestral lineage (for example, Betulaceae) (weighted linear regression: $r^2 = 0.54$; $P < 0.001$). Herbaceous plants evolved more recently (Fig. 1b, green circles) and—with the exception of Amaryllidaceae and Boraginaceae—are characterized by thin roots and SRLs that exceed those of woody plants (Fig. 1a). Together, these patterns broadly characterize an evolutionary transition from ancient tree taxa defined by thick first-order roots, to more recently radiated[20,21] woody and herbaceous plants with thin roots that can explore markedly greater lengths of soil per unit of carbon invested.

The trend towards thinner roots has had major consequences for the symbiosis between plant roots and mycorrhizal fungi. We found that mycorrhizal colonization (that is, the percentage of root length colonized) declines as roots get thinner (Fig. 1c) and that herbaceous roots have approximately 30% less colonization than woody plants at the same root diameter (linear model; $r^2 = 0.63$ with the difference between herbaceous and woody plants at $P < 0.001$). In addition, herbaceous plants have on average 33% lower root tissue density than woody plants (Fig. 1d; unequal variance t-test; $P < 0.001$), though considerable unexplained variation exists across taxa. These differences suggest that first-order roots have become less dependent on mycorrhizae as they have evolved thinner diameters, and that the innovation of the short-lived herbaceous growth form has fundamentally changed the relationship between root diameter and mycorrhizal colonization.

A phylogenetic independence contrasts analysis[17] confirmed that variation in root diameter, SRL and mycorrhizal colonization is strongly influenced by evolutionary history (Blomberg's K value in Extended Data Table1). By contrast, root chemical traits did not display a clear phylogenetic signal, which indicates that for these traits ecological variation overshadows evolutionary constraints[22].

Taken together, our results identify a general evolutionary trend in vascular plants from thick roots that rely on mycorrhizal fungi for resource acquisition to thin roots that can explore the soil at high carbon-use efficiency, but with less reliance on mycorrhizae. The observed root trait combinations imply selection for two contrasting plant strategies: (i) a conservative strategy in which carbon allocation to mycorrhizae enhances the ability of plants to compete in environments with stable resources and intense plant–plant competition; and (ii) an opportunistic strategy in which thin roots benefit plants in less predictable environments (for example, seasonal drought or cold), where rapid root growth response to a fluctuating resource supply is rewarded.

It is less clear, however, why herbaceous plants have lower mycorrhizal colonization than woody plants at similar root diameters (Fig. 1c), although the softer tissues of herbaceous plants may cause their roots to be less permanent than those of woody plants (Extended Data Fig. 4c) and therefore less able to maintain stable mycorrhizal relationships.

We next evaluated whether the distribution of root diameter changed across biomes that may differ in the pattern and stability of their resource supplies. First, we found an overall trend of decreasing variance in root diameter of woody plants from the more stable conditions of tropical and sub-tropical forests to the highly seasonal conditions of boreal and desert biomes (Fig. 2, Levene's test, Extended Data Fig. 5c). Second, woody plants were limited to thin-rooted species in the most seasonal biomes, but the diameter of herbaceous plant roots did not differ systematically across biomes (Extended Data Fig. 5a, b).

These patterns are consistent with biome-specific differences in both evolutionary history and the stability of resource supply and abiotic conditions. The tropical forest biome is evolutionarily ancient[23], characterized by seasonally stable supplies of soil resources and holds species that range from ancestral thick-rooted to more-derived thin-rooted taxa. By contrast, boreal and desert biomes are evolutionarily young[24] and have been colonized mainly by thin-rooted species that, in theory, can rapidly respond[25] to fluctuating soil resources and seasonally inhospitable conditions. The coexistence in the tropical biome of thin-rooted plants and plants pursuing more-ancient thick-rooted strategies suggests that the heterogeneity of this biome is sufficient to maintain a range of niche conditions for diverse belowground strategies.

Our findings suggest that at the timescale of plant evolution innovations of belowground traits have been important for preparing plants to colonize new habitats, and for the rich generation of biodiversity within and across biomes. The dominant dimension of trait evolution for first-order roots has been a decrease in diameter, which has reduced dependence on mycorrhizal fungi, increased the efficiency of root growth and thus enhanced the ability of plants to leverage photosynthetic carbon for soil exploration. An improved functional understanding of root traits is critical for comprehending the history and distribution of plant life, and may help to predict the risk of species extinction and to conserve biodiversity in the face of environmental change.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 August 2017; accepted 24 January 2018.
Published online 21 February 2018.

1. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389,** 33–39 (1997).
2. Field, K. J., Pressel, S., Duckett, J. G., Rimington, W. R. & Bidartondo, M. I. Symbiotic options for the conquest of land. *Trends Ecol. Evol.* **30,** 477–486 (2015).
3. Bardgett, R. D., Mommer, L. & De Vries, F. T. Going underground: root traits as drivers of ecosystem processes. *Trends Ecol. Evol.* **29,** 692–699 (2014).
4. Weemstra, M. *et al.* Towards a multidimensional root trait framework: a tree root review. *New Phytol.* **211,** 1159–1169 (2016).
5. Díaz, S. *et al.* The global spectrum of plant form and function. *Nature* **529,** 167–171 (2016).
6. Wright, I. J. *et al.* The worldwide leaf economics spectrum. *Nature* **428,** 821–827 (2004).
7. Roumet, C. *et al.* Root structure–function relationships in 74 species: evidence of a root economics spectrum related to carbon economy. *New Phytol.* **210,** 815–826 (2016).

8. Pregitzer, K. S. *et al.* Fine root architecture of nine North American trees. *Ecol. Monogr.* **72,** 293–309 (2002).
9. McCormack, M. L. *et al.* Redefining fine roots improves understanding of below-ground contributions to terrestrial biosphere processes. *New Phytol.* **207,** 505–518 (2015).
10. Chen, W. *et al.* Root morphology and mycorrhizal symbioses together shape nutrient foraging strategies of temperate trees. *Proc. Natl Acad. Sci. USA* **113,** 8741–8746 (2016).
11. Norby, R. J., Ledford, J., Reilly, C. D., Miller, N. E. & O'Neill, E. G. Fine-root production dominates response of a deciduous forest to atmospheric CO_2 enrichment. *Proc. Natl Acad. Sci. USA* **101,** 9689–9693 (2004).
12. Jackson, R. B., Mooney, H. A. & Schulze, E. D. A global budget for fine root biomass, surface area, and nutrient contents. *Proc. Natl Acad. Sci. USA* **94,** 7362–7366 (1997).
13. McCormack, M. L. *et al.* Building a better foundation: improving root-trait measurements to understand and model plant and ecosystem processes. *New Phytol.* **215,** 27–37 (2017).
14. Chen, W. L., Zeng, H., Eissenstat, D. M. & Guo, D. Variation of first-order root traits across climatic gradients and evolutionary trends in geological time. *Glob. Ecol. Biogeogr*. **22,** 846–856 (2013).
15. Guo, D. *et al.* Anatomical traits associated with absorption and mycorrhizal colonization are linked to root branch order in twenty-three Chinese temperate tree species. *New Phytol*. **180,** 673–683 (2008).
16. Reich, P. B. *et al.* Scaling of respiration to nitrogen in leaves, stems and roots of higher land plants. *Ecol. Lett.* **11,** 793–801 (2008).
17. Valverde-Barrantes, O. J., Freschet, G. T., Roumet, C. & Blackwood, C. B. A worldview of root traits: the influence of ancestry, growth form, climate and mycorrhizal association on the functional trait variation of fine-root tissues in seed plants. *New Phytol.* **215,** 1562–1573 (2017).
18. Iversen, C. M. *et al.* A global fine-root ecology database to address below-ground challenges in plant ecology. *New Phytol.* **215,** 15–26 (2017).
19. Kong, D. *et al.* Leading dimensions in absorptive root trait variation across 96 subtropical forest species. *New Phytol.* **203,** 863–872 (2014).
20. Feild, T. S. & Arens, N. C. Form, function and environments of the early angiosperms: merging extant phylogeny and ecophysiology with fossils. *New Phytol*. **166,** 383–408 (2005).
21. Comas, L. H. *et al.* Evolutionary patterns and biogeochemical significance of angiosperm root traits. *Int. J. Plant Sci.* **173,** 584–595 (2012).
22. Ackerly, D. D. & Reich, P. B. Convergence and correlations among leaf size and function in seed plants: a comparative test using independent contrasts. *Am. J. Bot.* **86,** 1272–1281 (1999).
23. Wing, S. L. *et al.* Late Paleocene fossils from the Cerrejon Formation, Colombia, are the earliest record of Neotropical rainforest. *Proc. Natl Acad. Sci. USA* **106,** 18627–18632 (2009).
24. Woodward, F. I., Lomas, M. R. & Kelly, C. K. Global climate and the distribution of plant biomes. *Philos. Trans. Roy. Soc. Lond. B.* **359,** 1465–1476 (2004).
25. Eissenstat, D. M., Kucharski, J. M., Zadworny, M., Adams, T. S. & Koide, R. T. Linking root traits to nutrient foraging in arbuscular mycorrhizal trees in a temperate forest. *New Phytol.* **208,** 114–124 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Ma, L. Li, Y. Yue, M. Liu, F. Ma, H. Li, D. Kong, B. Liu and K. Sun for collecting data; X. Liu and X. Deng for their assistance in field sampling; and all members of field research stations of the Chinese Academy of Sciences for their support. This study was funded by the Natural Science Foundation of China (NSFC Grants 31325006, 31530011, and 41571130041).

Author Contributions Z.M., D.G. and L.O.H. developed the overall conceptual approach and analysis. Z.M. compiled and analysed the data, and X.X. provided nitrogen-uptake-rate data. R.D.B., D.M.E. and M.L.M. contributed to the formulation of research questions and interpretations. Z.M., D.G., M.L. and L.O.H. wrote the paper and all authors contributed to revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to L.O.H. (lhedin@princeton.edu).

METHODS

Sampling approach. We collected roots from natural plant communities across seven major biomes and three continents (Asia, Europe and North America) between 2004 and 2016. Our sampling sites range from −1.4 °C to 22.4 °C in mean annual temperature, and from 35 mm to 2,651 mm in mean annual precipitation. At each site, we selected common indigenous species that are representative of the local plant community. For each species, we sampled multiple root branches or segments from at least three individual plants to derive the mean species trait value. For species that occupied more than one sample location, we merged the local means into one species trait value. Eleven species occurred in more than a single biome; for these we calculated a mean value for each biome.

In mixed-species ecosystems we identified roots to the species level by tracing a root to its parent tree. During the growing season, we selected mature individuals and excavated the surface soil (0–20 cm) around the plant stem to expose lateral roots. We then sampled multiple intact root branches and gently cleared the attached soil. Sampled roots were bagged and immediately placed in a cooler, and then either transferred to a refrigerator for processing within the next few days or kept frozen until later laboratory analyses.

Laboratory analyses of root functional traits. We adopted a previously established (for details, see Supplementary Information, note 2, and ref. 8) approach based on root branching-order, in which absorptive fine roots are sorted on the basis of their position in the branching architecture. We dissected root branches according to standard methodology[15] and analysed four morphological traits (diameter, SRL, root tissue density and root length), three physiological–chemical traits (root nitrogen content, root carbon content and root carbon-to-nitrogen ratio) and the extent of mycorrhizal colonization (Supplementary Information, note 1).

Root diameter and length were measured using a stereomicroscope with an ocular micrometer. SRL was determined by dividing root length by the dry biomass weight. We calculated the volume of root segments from root diameter and length, assuming segments are cylinders. Root tissue density was then calculated using dry mass and volume. Sampled roots were dried in an oven at 60 °C for 48 h, and then ground to fine powder with a ball-mill for subsequent measurements of carbon and nitrogen on an elemental analyser (Vario EL Cube; Elementar).

We measured the length of root colonized by mycorrhizal fungi in 137 species from sub-tropical[19] forests, temperate forests[15] and temperate grasslands[26]. We calculated the percentage of mycorrhizal colonization by sampling first-order roots, and determining by microscope the presence of either arbuscular mycorrhizal or ectomycorrhizal fungal structures within an individual root segment. For arbuscular mycorrhizal fungi we used a standard staining technique to identify coils and arbuscules[19,26]; no stain was needed to identify ectomycorrhizal fungal sheaths[19]. For each individual plant, we selected at least ten root branches (containing multiple orders of roots). For each species, the percentage of colonization was calculated across 20–150 randomly selected first-order root segments[19,26] to ensure that each segment length was consistent across all roots sampled. The percentage of length colonization was calculated as the ratio of the sum of infected root segments over all root segments examined. We used two different techniques: one based on cross-section analysis (MC_1, denoting mycorrhizal colonization technique 1; $n = 110$ species) and one based on scanning the root surface (MC_2, denoting mycorrhizal colonization technique 2; $n = 27$ species); both enabled us to quantify fungal association within a standardized root area. We kept the effective area examined per root segment approximately the same for both methods (173 versus 169 mm^2), such that the results are equivalent ($MC_2 = 1.02 \times MC_1 - 0.02$; $r^2 = 0.988$).

Because distal fine roots (that is, first-order roots) are primarily responsible for plant nutrient acquisition[8,9], we focused our analyses on first-order root traits. We accumulated 480 species-specific observations, of which 187 are unpublished and 256 have previously been published[14,15,19,25,26]. To enhance the coverage of some biomes (for example, boreal and Mediterranean), we added 37 previously published[8,27–29] observations into our dataset (comprising ~5% of the final dataset), taking care to include only studies of first-order roots and consistent methods. In total, we gathered traits from 369 species (281 woody, and 88 herbaceous, species; Extended Data Fig. 6 and Extended Data Table 1) covering 210 genera and 79 families.

Root lifespan. We collected data on plant root lifespan for 40 species using *in situ* minirhizotrons in boreal and temperate forests of Europe[30], Asia[31,32] and North America[33,34], with individual measures spanning at least one year. We acquired additional previously published lifespan data, selecting only studies of distal roots using *in situ* minirhizotrons or root windows[35–56]. When corresponding root traits (for example, diameter or SRL) were not available, we used species-specific observations from our own dataset to match the lifespan data. In total, we obtained 70 species-specific observations and 13 community observations across 5 biomes.

Root nitrogen uptake rates. We measured per-biomass root nitrogen uptake rates in 34 plant species using 2 standard approaches: (i) by isolating an intact living root branch and exposing it to a hydroponic solution with isotopically labelled ammonium nitrate (intrusive approach, elevated nitrogen concentration; see ref. 57); and (ii) by applying nutrient solution to soil and allowing plant roots to take up nutrients *in situ* (non-intrusive, low nitrogen concentration; see ref. 58). The first approach enables an estimation of the maximum uptake rate of absorptive roots and the second approach more accurately reflects the uptake rate of roots in natural conditions; we analysed the resulting two datasets separately. We acquired previously published data for an additional 107 species[59–101]. The resulting dataset is summarized in Extended Data Fig. 2d.

Species and phylogeny. We conducted phylogenetic analyses of our species from 210 genera and 79 families, confirming species names using The Plant List (http://www.theplantlist.org). We followed the APG III phylogenetic system[102] in all analyses and used PHYLOCOM[103] to construct phylogenetic trees (Extended Data Fig. 7). Following a previous analysis[104], we defined the divergence time of a plant family using the earliest diverging genus within that family. We calculated Blomberg's K statistic[105] using the 'Picante' package in R and evaluated the strength of the phylogenetic signal for each trait; a large Blomberg's K value is thought to indicate phylogenetic conservatism. We performed phylogenetic independent contrasts analyses to correct for shared evolutionary histories among traits and to look for the effect of environmental influences (Extended Data Table 3).

Quantitative and statistical analyses. Shapiro–Wilk tests revealed that all of our traits were significantly non-normal ($P < 0.05$), which we corrected by $\log_{10}$-transforming our data. We derived the nonlinear relationship between SRL and root diameter based on inherent biophysical constraints, as discussed in Supplementary Information, note 1. We fit the equation using the average root tissue density across all species and evaluated the sensitivity to variation in root tissue density across a tenfold range.

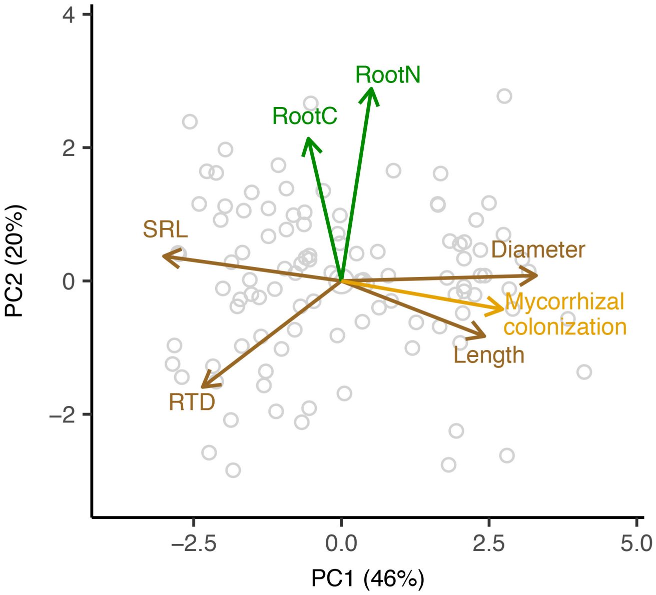
All statistical analyses were performed using the R software (version 2.15.0.), using the 'Factor R' package for principal component analyses and the 'lme4' package for mixed linear effect models. The linear regression between root diameter and divergence time was weighted by the number of species within each family. We used linear regression to test the effect of root diameter and growth form (woody versus herbaceous) on root mycorrhizal colonization. Across biomes, we examined equality of variance in root diameter using Levene's test and differences in mean root diameter using a linear mixed effects model.

Code availability. The R scripts used in Figs 1 and 2 are available from the corresponding author upon reasonable request.

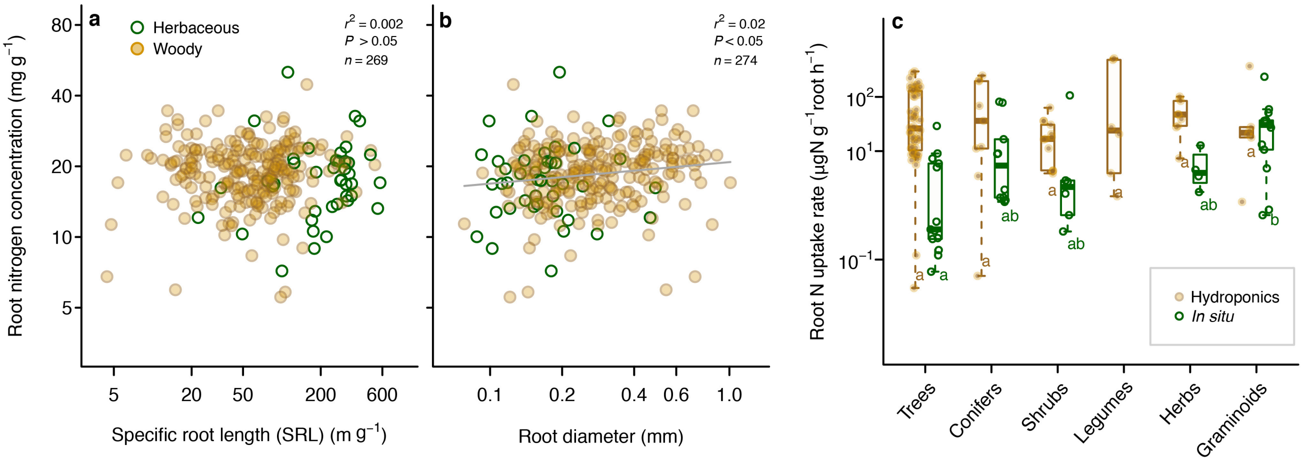
Data availability. The data that support the findings of this study are available from the corresponding authors upon reasonable request.

26. Li, H., Liu, B., McCormack, M. L., Ma, Z. & Guo, D. Diverse belowground resource strategies underlie plant species coexistence and spatial distribution in three grasslands along a precipitation gradient. *New Phytol.* **216,** 1140–1150 (2017).
27. Hummel, I. *et al.* Relating root structure and anatomy to whole-plant functioning in 14 herbaceous Mediterranean species. *New Phytol*. **173,** 313–321 (2007).
28. Ostonen, I. *et al.* Fine root foraging strategies in Norway spruce forests across a European climate gradient. Glob. Chang. *Biol.* **17,** 3620–3632 (2011).
29. Leppälammi-Kujansuu, J. *et al.* Fine root longevity and carbon input into soil from below- and aboveground litter in climatically contrasting forests. *For. Ecol. Manage.* **326,** 79–90 (2014).
30. Withington, J. M., Reich, P. B., Oleksyn, J. & Eissenstat, D. M. Comparisons of structure and life span in roots and leaves among temperate trees. *Ecol. Monogr.* **76,** 381–397 (2006).
31. Sun, K., McCormack, M. L., Li, L., Ma, Z. & Guo, D. Fast-cycling unit of root turnover in perennial herbaceous plants in a cold temperate ecosystem. *Sci. Rep.* **6,** 19698 (2016).
32. Xia, M., Guo, D. & Pregitzer, K. S. Ephemeral root modules in *Fraxinus mandshurica*. *New Phytol*. **188,** 1065–1074 (2010).
33. McCormack, M. L., Adams, T. S., Smithwick, E. A. H. & Eissenstat, D. M. Predicting fine root lifespan from plant functional traits in temperate trees. *New Phytol*. **195,** 823–831 (2012).
34. Valenzuela-Estrada, L. R., Vera-Caraballo, V., Ruth, L. E. & Eissenstat, D. M. Root anatomy, morphology, and longevity among root orders in *Vaccinium corymbosum* (Ericaceae). *Am. J. Bot.* **95,** 1506–1514 (2008).
35. Liu, B., He, J., Zeng, F., Lei, J. & Arndt, S. K. Life span and structure of ephemeral root modules of different functional groups from a desert system. *New Phytol.* **211,** 103–112 (2016).
36. Hansson, K., Helmisaari, H.-S., Sah, S. P. & Lange, H. Fine root production and turnover of tree and understorey vegetation in Scots pine, silver birch and Norway spruce stands in SW Sweden. *For. Ecol. Manage.* **309,** 58–65 (2013).
37. Huang, J. X. *et al.* Fine root longevity and controlling factors in subtropical *Altingia grlilipes* and *Castanopsis carlesii* forests. *Acta Ecol. Sin.* **32,** 1932–1942 (2012).
38. Yavitt, J. B., Harms, K. E., Garcia, M. N., Mirabello, M. J. & Wright, S. J. Soil fertility and fine root dynamics in response to 4 years of nutrient (N, P, K) fertilization in a lowland tropical moist forest, Panama. *Austral. Ecol.* **36,** 433–445 (2011).
39. Ling, H. *et al.* Influencing factors of fine root lifespans in two Chinese fir plantations in subtropical China. *Acta Ecol. Sin.* **31,** 1130–1138 (2011).

40. Stover, D. B., Day, F. P., Drake, B. G. & Hinkle, C. R. The long-term effects of CO_2 enrichment on fine root productivity, mortality, and survivorship in a scrub-oak ecosystem at Kennedy Space Center, Florida, USA. *Environ. Exp. Bot.* **69,** 214–222 (2010).
41. Krasowski, M. J., Lavigne, M. B., Olesinski, J. & Bernier, P. Y. Advantages of long-term measurement of fine root demographics with a minirhizotron at two balsam fir sites. *Can. J. For. Res.* **40,** 1128–1135 (2010).
42. Huang, G., Zhao, X., Zhao, H., Huang, Y. & Zuo, X. Linking root morphology, longevity and function to root branch order: a case study in three shrubs. *Plant Soil* **336,** 197–208 (2010).
43. Girardin, C. A. J. *et al.* Net primary productivity allocation and cycling of carbon along a tropical forest elevational transect in the Peruvian Andes. Glob. Chang. *Biol.* **16,** 3176–3192 (2010).
44. Espeleta, J. F., West, J. B. & Donovan, L. A. Tree species fine-root demography parallels habitat specialization across a sandhill soil resource gradient. *Ecology* **90,** 1773–1787 (2009).
45. Vargas, R. & Allen, M. F. Dynamics of fine root, fungal rhizomorphs, and soil respiration in a mixed temperate forest: integrating sensors and observations. *Vadose Zone J.* **7,** 1055–1064 (2008).
46. Graefe, S., Hertel, D. & Leuschner, C. Fine root dynamics along a 2,000-m elevation transect in South Ecuadorian mountain rainforests. *Plant Soil* **313,** 155–166 (2008).
47. Green, J. J., Dawson, L. A., Proctor, J., Duff, E. I. & Elston, D. A. Fine root dynamics in a tropical rain forest is influenced by rainfall. *Plant Soil* **276,** 23–32 (2005).
48. Baddeley, J. A. & Watson, C. A. Influences of root diameter, tree age, soil depth and season on fine root survivorship in *Prunus avium*. *Plant Soil* **276,** 15–22 (2005).
49. Kern, C. C., Friend, A. L., Johnson, J. M. F. & Coleman, M. D. Fine root dynamics in a developing *Populus deltoides* plantation. *Tree Physiol.* **24,** 651–660 (2004).
50. Wells, C. E., Glenn, D. M. & Eissenstat, D. M. Changes in the risk of fine-root mortality with age: a case study in peach, *Prunus persica* (Rosaceae). *Am. J. Bot.* **89,** 79–87 (2002).
51. King, J. S. *et al.* Below-ground carbon input to soil is controlled by nutrient availability and fine root dynamics in loblolly pine. *New Phytol.* **154,** 389–398 (2002).
52. Gill, R. A., Burke, I. C., Lauenroth, W. K. & Milchunas, D. G. Longevity and turnover of roots in the shortgrass steppe: influence of diameter and depth. *Plant Ecol.* **159,** 241–251 (2002).
53. Wells, C. E. & Eissenstat, D. M. Marked differences in survivorship among apple roots of different diameters. *Ecology* **82,** 882–892 (2001).
54. Tierney, G. L. & Fahey, T. J. Evaluating minirhizotron estimates of fine root longevity and production in the forest floor of a temperate broadleaf forest. *Plant Soil* **229,** 167–176 (2001).
55. Majdi, H., Damm, E. & Nylund, J.-E. Longevity of mycorrhizal roots depends on branching order and nutrient availability. *New Phytol.* **150,** 195–202 (2001).
56. Coleman, M. D., Dickson, R. E. & Isebrands, J. G. Contrasting fine-root production, survival and soil CO_2 efflux in pine and poplar plantations. *Plant Soil* **225,** 129–139 (2000).
57. Chapin, F. S. III, Moilanen, L. & Kielland, K. Preferential use of organic nitrogen for growth by a nonmycorrhizal Arctic sedge. *Nature* **361,** 150–153 (1993).
58. McKane, R. B. *et al.* Resource-based niches provide a basis for plant species diversity and dominance in arctic tundra. *Nature* **415,** 68–71 (2002).
59. Aanderud, Z. T. & Bledsoe, C. S. Preferences for ^{15}N-ammonium, ^{15}N-nitrate, and ^{15}N-glycine differ among dominant exotic and subordinate native grasses from a California oak woodland. *Environ. Exp. Bot.* **65,** 205–209 (2009).
60. Andersen, K. M. & Turner, B. L. Preferences or plasticity in nitrogen acquisition by understorey palms in a tropical montane forest. *J. Ecol.* **101,** 819–825 (2013).
61. Andresen, L. C., Michelsen, A., Jonasson, S. & Ström, L. Seasonal changes in nitrogen availability, and root and microbial uptake of $^{15}N^{13}C_9$-phenylalanine and ^{15}N-ammonium *in situ* at a temperate heath. *Appl. Soil Ecol.* **51,** 94–101 (2011).
62. Averill, C. & Finzi, A. Increasing plant use of organic nitrogen with elevation is reflected in nitrogen uptake rates and ecosystem δ^{15}N. *Ecology* **92,** 883–891 (2011).
63. Bardgett, R. D., Streeter, T. C. & Bol, R. Soil microbes compete effectively with plants for organic-nitrogen inputs to temperate grasslands. *Ecology* **84,** 1277–1287 (2003).
64. Boczulak, S. A., Hawkins, B. J. & Roy, R. Temperature effects on nitrogen form uptake by seedling roots of three contrasting conifers. *Tree Physiol.* **34,** 513–523 (2014).
65. Cheng, X. & Bledsoe, C. S. Competition for inorganic and organic N by blue oak (*Quercus douglasii*) seedlings, an annual grass, and soil microorganisms in a pot study. *Soil Biol. Biochem.* **36,** 135–144 (2004).
66. Dunn, R. M., Mikola, J., Bol, R. & Bardgett, R. D. Influence of microbial activity on plant–microbial competition for organic and inorganic nitrogen. *Plant Soil* **289,** 321–334 (2006).
67. Finzi, A. C. & Berthrong, S. T. The uptake of amino acids by microbes and trees in three cold-temperate forests. *Ecology* **86,** 3345–3353 (2005).
68. Gallet-Budynek, A. *et al.* Intact amino acid uptake by northern hardwood and conifer trees. *Oecologia* **160,** 129–138 (2009).
69. Harrison, K. A., Bol, R. & Bardgett, R. D. Preferences for different nitrogen forms by coexisting plant species and soil microbes. *Ecology* **88,** 989–999 (2007).
70. Henry, H. A. L. & Jefferies, R. L. Interactions in the uptake of amino acids, ammonium and nitrate ions in the Arctic salt-marsh grass, *Puccinellia phryganodes*. *Plant Cell Environ.* **26,** 419–428 (2003).
71. Jin, V. L. & Evans, R. D. Microbial ^{13}C utilization patterns via stable isotope probing of phospholipid biomarkers in Mojave Desert soils exposed to ambient and elevated atmospheric CO_2. *Glob. Chang. Biol.* **16,** 2334–2344 (2010).
72. Jin, V. L., Romanek, C. S., Donovan, L. A. & Sharitz, R. R. Soil nitrogen availability and *in situ* nitrogen uptake by *Acer rubrum* L. and *Pinus palustris* Mill. in the southeastern U. S. Coastal Plain. *J. Torrey Bot. Soc.* **137,** 339–347 (2010).
73. Kahmen, A., Livesley, S. J. & Arndt, S. K. High potential, but low actual, glycine uptake of dominant plant species in three Australian land-use types with intermediate N availability. *Plant Soil* **325,** 109–121 (2009).
74. Kaštovská, E. & Šantrůčková, H. Comparison of uptake of different N forms by soil microorganisms and two wet-grassland plants: a pot study. *Soil Biol. Biochem.* **43,** 1285–1291 (2011).
75. Li, C. *et al.* Inorganic and organic nitrogen uptake by nine dominant subtropical tree species. *iForest (Viterbo)* **9,** 253–258 (2015).
76. McFarland, J. W. *et al.* Cross-ecosystem comparisons of *in situ* plant uptake of amino acid-N and NH_4^+. *Ecosystems* **13,** 177–193 (2010).
77. Metcalfe, R. J., Nault, J. & Hawkins, B. J. Adaptations to nitrogen form: comparing inorganic nitrogen and amino acid availability and uptake by four temperate forest plants. *Can. J. For. Res.* **41,** 1626–1637 (2011).
78. Mozdzer, T. J., Zieman, J. C. & McGlathery, K. J. Nitrogen uptake by native and invasive temperate coastal macrophytes: importance of dissolved organic nitrogen. *Estuaries Coast.* **33,** 784–797 (2010).
79. Nie, M. *et al.* Plants' use of different nitrogen forms in response to crude oil contamination. *Environ. Pollut.* **159,** 157–163 (2011).
80. Nordin, A., Högberg, P. & Näsholm, T. Soil nitrogen form and plant nitrogen uptake along a boreal forest productivity gradient. *Oecologia* **129,** 125–132 (2001).
81. Öhlund, J. & Näsholm, T. Growth of conifer seedlings on organic and inorganic nitrogen sources. *Tree Physiol.* **21,** 1319–1326 (2001).
82. Ouyang, S. *et al.* Nitrogen competition between three dominant plant species and microbes in a temperate grassland. *Plant Soil* **408,** 121–132 (2016).
83. Paulding, E. M., Baker, A. J. M. & Warren, C. R. Competition for nitrogen by three sympatric species of *Eucalyptus*. *Ann. For. Sci.* **67,** 406 (2010).
84. Persson, J. *et al.* Nitrogen acquisition from inorganic and organic sources by boreal forest plants in the field. *Oecologia* **137,** 252–257 (2003).
85. Persson, J. & Näsholm, T. Regulation of amino acid uptake in conifers by exogenous and endogenous nitrogen. *Planta* **215,** 639–644 (2002).
86. Pfautsch, S., Rennenberg, H., Bell, T. L. & Adams, M. A. Nitrogen uptake by *Eucalyptus regnans* and *Acacia* spp. – preferences, resource overlap and energetic costs. *Tree Physiol.* **29,** 389–399 (2009).
87. Rains, K. C. & Bledsoe, C. S. Rapid uptake of ^{15}N-ammonium and glycine-^{13}C, ^{15}N by arbuscular and ericoid mycorrhizal plants native to a Northern California coastal pygmy forest. *Soil Biol. Biochem.* **39,** 1078–1086 (2007).
88. Schmidt, S. & Stewart, G. R. Waterlogging and fire impacts on nitrogen availability and utilization in a subtropical wet heathland (wallum). *Plant Cell Environ.* **20,** 1231–1241 (1997).
89. Schmidt, S. & Stewart, G. R. Glycine metabolism by plant roots and its occurrence in Australian plant communities. *Aust. J. Plant Physiol.* **26,** 253–264 (1999).
90. Scott, E. E. & Rothstein, D. E. Amino acid uptake by temperate tree species characteristic of low- and high-fertility habitats. *Oecologia* **167,** 547–557 (2011).
91. Simon, J. *et al.* Competition for nitrogen between adult European beech and its offspring is reduced by avoidance strategy. *For. Ecol. Manage.* **262,** 105–114 (2011).
92. Simon, J., Waldhecker, P., Brüggemann, N. & Rennenberg, H. Competition for nitrogen sources between European beech (*Fagus sylvatica*) and sycamore maple (*Acer pseudoplatanus*) seedlings. *Plant Biol.* **12,** 453–458 (2010).
93. Stoelken, G., Simon, J., Ehlting, B. & Rennenberg, H. The presence of amino acids affects inorganic N uptake in non-mycorrhizal seedlings of European beech (*Fagus sylvatica*). *Tree Physiol.* **30,** 1118–1128 (2010).
94. Wallander, H., Arnebrant, K., Östrand, F. & Kårén, O. Uptake of N^{15}-labelled alanine, ammonium and nitrate in *Pinus sylvestris* L. ectomycorrhiza growing in forest soil treated with nitrogen, sulphur or lime. *Plant Soil* **195,** 329–338 (1997).
95. Wanek, W., Arndt, S. K., Huber, W. & Popp, M. Nitrogen nutrition during ontogeny of hemiepiphytic *Clusia* species. *Funct. Plant Biol.* **29,** 733–740 (2002).
96. Warren, C. R. Potential organic and inorganic N uptake by six *Eucalyptus* species. *Funct. Plant Biol.* **33,** 653–660 (2006).
97. Warren, C. R. Does nitrogen concentration affect relative uptake rates of nitrate, ammonium, and glycine? *J. Plant Nutr. Soil Sci.* **172,** 224–229 (2009).
98. Warren, C. R. & Adams, P. R. Uptake of nitrate, ammonium and glycine by plants of Tasmanian wet eucalypt forests. *Tree Physiol.* **27,** 413–419 (2007).
99. Wei, L., Chen, C. & Yu, S. Uptake of organic nitrogen and preference for inorganic nitrogen by two Australian native Araucariaceae species. *Plant Ecol. Divers.* **8,** 259–264 (2015).
100. Weigelt, A., Bol, R. & Bardgett, R. D. Preferential uptake of soil nitrogen forms by grassland plant species. *Oecologia* **142,** 627–635 (2005).
101. Wu, J. *et al.* Mycorrhizas alter nitrogen acquisition by the terrestrial orchid *Cymbidium goeringii*. *Ann. Bot.* **111,** 1181–1187 (2013).
102. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* **161,** 105–121 (2009).
103. Webb, C. O., Ackerly, D. D. & Kembel, S. W. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* **24,** 2098–2100 (2008).
104. Wikström, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B* **268,** 2211–2220 (2001).
105. Blomberg, S. P., Garland, T., Jr & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57,** 717–745 (2003).



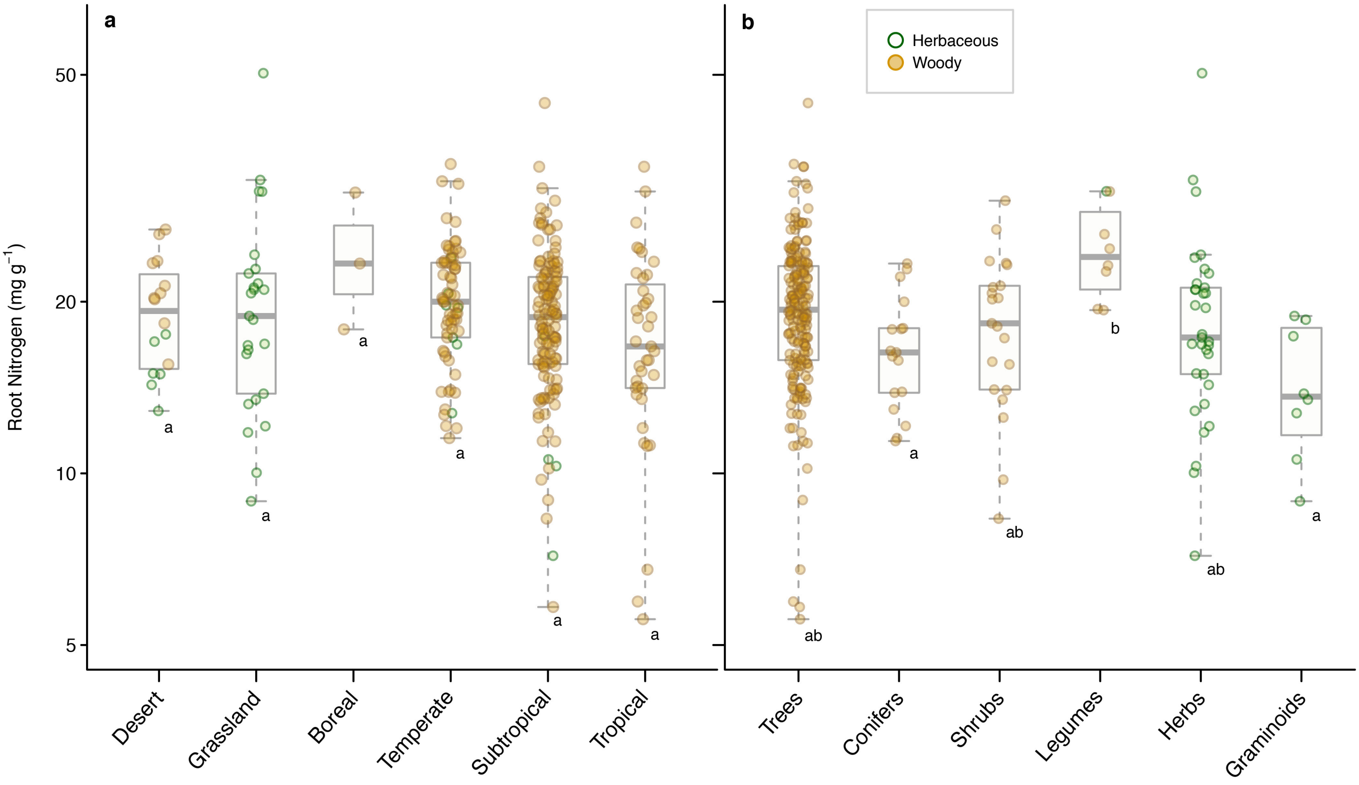
Extended Data Figure 1 | Principal component analysis of 7 root functional traits across 104 species. Trait loading on the plane defined by principal components 1 and 2 (PC1 and PC2). Brown arrows indicate four morphological traits; diameter, length, SRL and root tissue density (RTD). Green arrows indicate two physiological–chemical traits; root carbon (RootC) and root nitrogen (RootN). The yellow arrow shows mycorrhizal colonization. Three different analyses confirm the results shown here (detailed in Extended Data Table 2): (i) all data excluding the mycorrhizal colonization trait (*n* = 217 species); (ii) gaps in mycorrhizal colonization data interpolated using the regression from Fig. 1c (n = 217 species); and (iii) gaps in any trait value interpolated (*n* = 369) using the regressions in Fig. 1a, c or the multiple imputation method in the MICE R package).



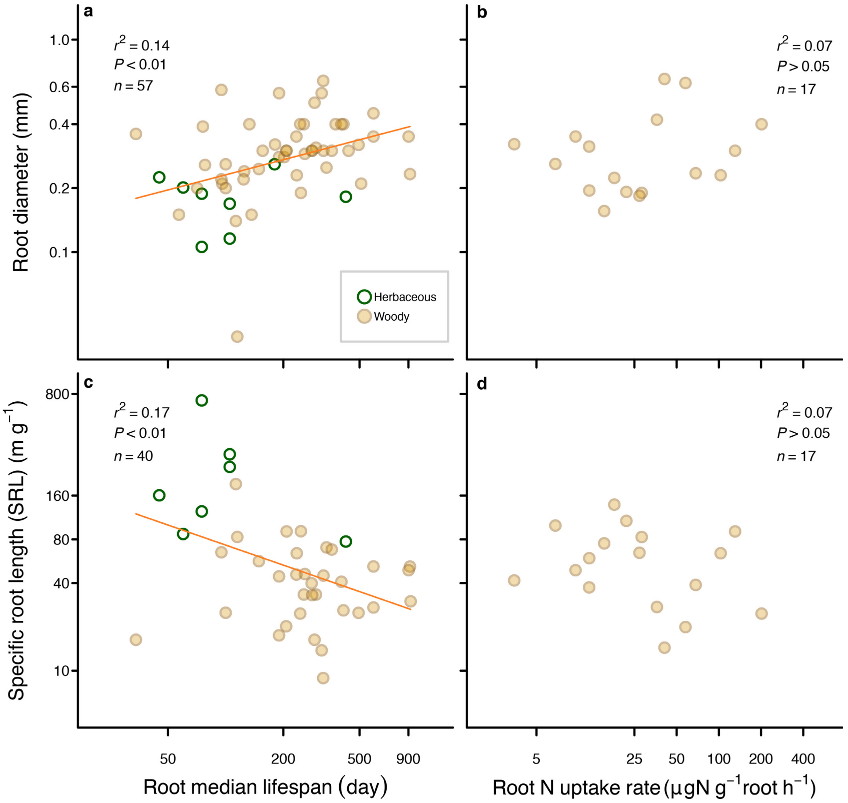
d

	Hydroponics* (μg N g^{-1} Root h^{-1})					*In situ*† (μg N g^{-1} Root h^{-1})				
Biome	*n*	Mean	SE	Minimum	Maximum	*n*	Mean	SE	Minimum	Maximum
Grassland	3	139.8	114.3	23.3	368.5	4	20.16	6.88	9.24	40.22
Boreal	5	199.24	42.57	74.44	336.67	17	24.72	11.30	0.33	164.08
Temperate	41	96.46	15.30	0.03	315.93	38	17.70	7.74	0.72	222.13
Subtropical	38	59.68	17.88	1.16	511.00	7	59.89	29.60	8.87	234.67
Tropical	24	29.47	8.03	5.61	193.85	9	1.64	1.39	0.06	12.73

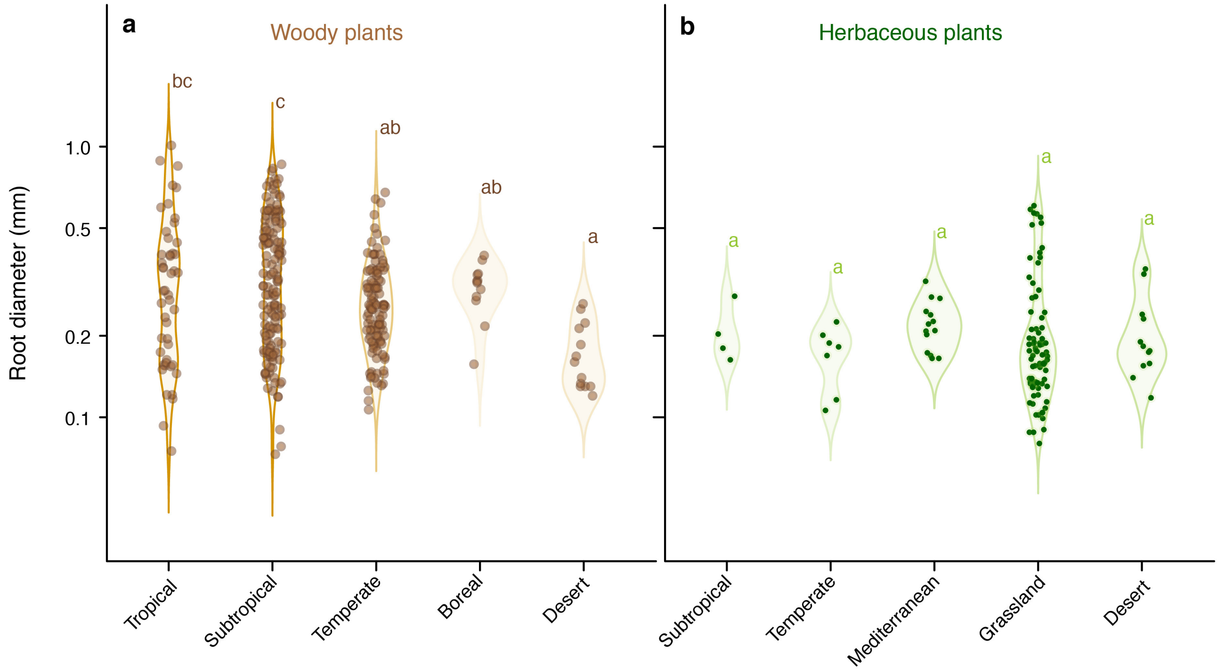
Extended Data Figure 2 | Root nitrogen concentration and root nitrogen uptake rate. a, There is no correlation between root nitrogen and SRL ($r^2 = 0.002$, $P = 0.81$, $n = 269$). **b**, There is no correlation between root nitrogen and diameter ($r^2 = 0.02$, $P = 0.01$, $n = 274$). Each point represents one species: brown, woody plants; green, herbaceous plants (**a**, **b**). **c**, Across plant growth forms, nitrogen uptake rates per root biomass did not vary significantly ($P > 0.05$) based on hydroponic measurements (brown). For *in situ* experiments (green) we observed a growth-form effect ($P < 0.01$), which was caused solely by higher root uptake in graminoid species compared to trees. All other growth forms were statistically indistinguishable. The letters 'a' and 'b' indicate significant difference based on an ANOVA across growth forms. **d**, Summary table of nitrogen uptake rates across different biomes by two approaches. This dataset included previously unpublished data from 22 species. A detailed description of these two approaches can be found in the Methods. Additional data were collected from refs 59–101.



Extended Data Figure 3 | First-order-root nitrogen concentration across biomes and plant functional groups. a, We did not detect a distinct pattern in first-order-root nitrogen concentration across biomes (**a**) (ANOVA; $P > 0.14$, $n = 284$ species). **b**, We detected a slight difference in first-order-root nitrogen concentration among plant functional groups (**b**) (ANOVA; $P < 0.01$, $n = 284$ species), which was mostly driven by the higher root nitrogen concentrations found in legumes. Each point represents one species; brown, woody plants; green, herbaceous plants. The letters 'a', 'b' and 'ab' indicate significant differences between categories.



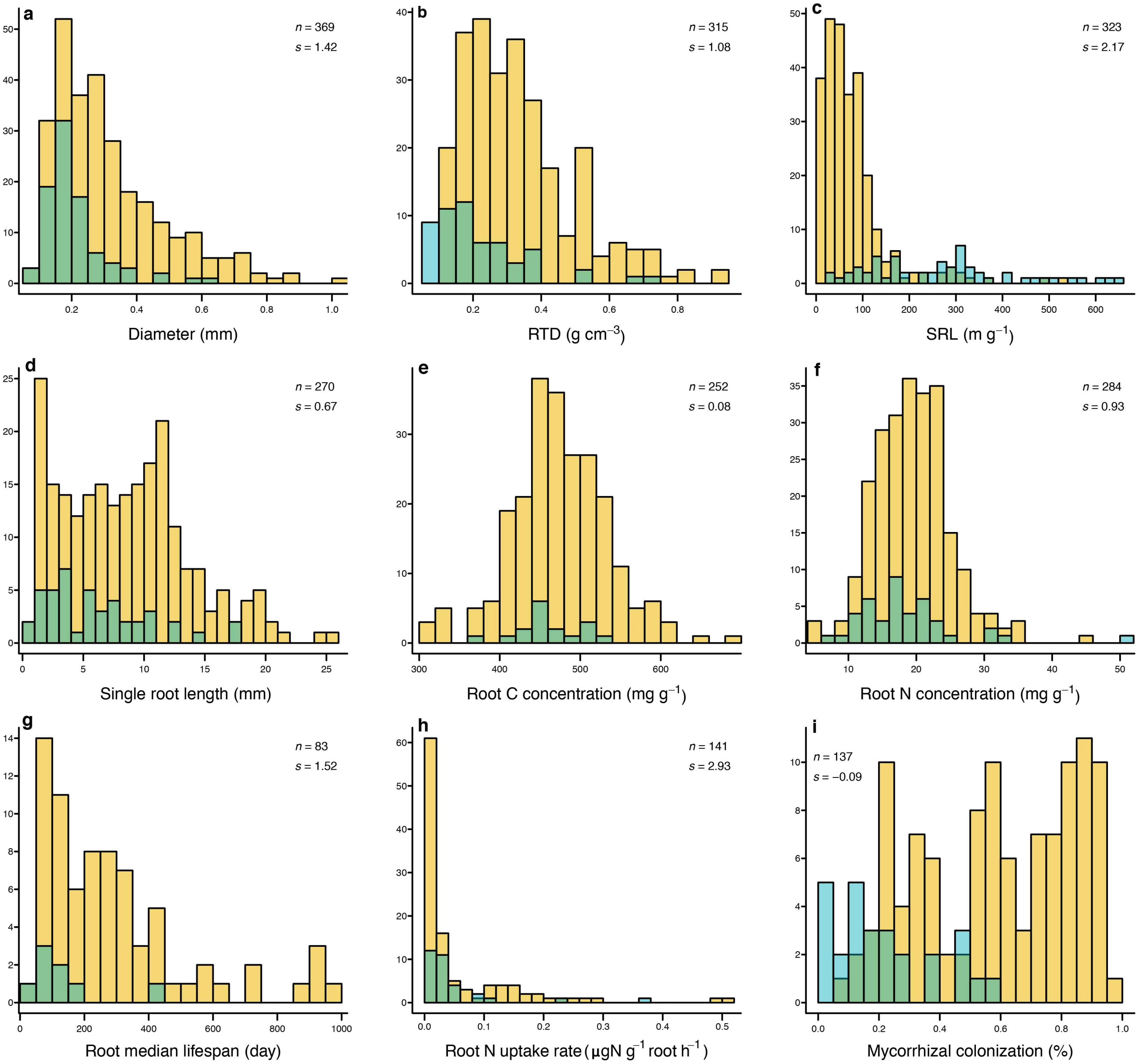
Extended Data Figure 4 | Relationship between root functional and morphological traits. a, **c**, Root median lifespan is significantly correlated with root diameter (**a**, $r^2 = 0.14$, $P < 0.01$, linear regression) and SRL (**c**, $r^2 = 0.17$, $P < 0.01$, linear regression). **b**, **d**, Root nitrogen uptake rate is not correlated with root diameter (**b**, $r^2 = 0.07$, $P > 0.05$, linear regression) or with SRL (**d**, $r^2 = 0.07$, $P = 0.29$, linear regression) in woody plants. Data are presented on a logarithmic scale ($\log_{10}$), with each point representing one species.



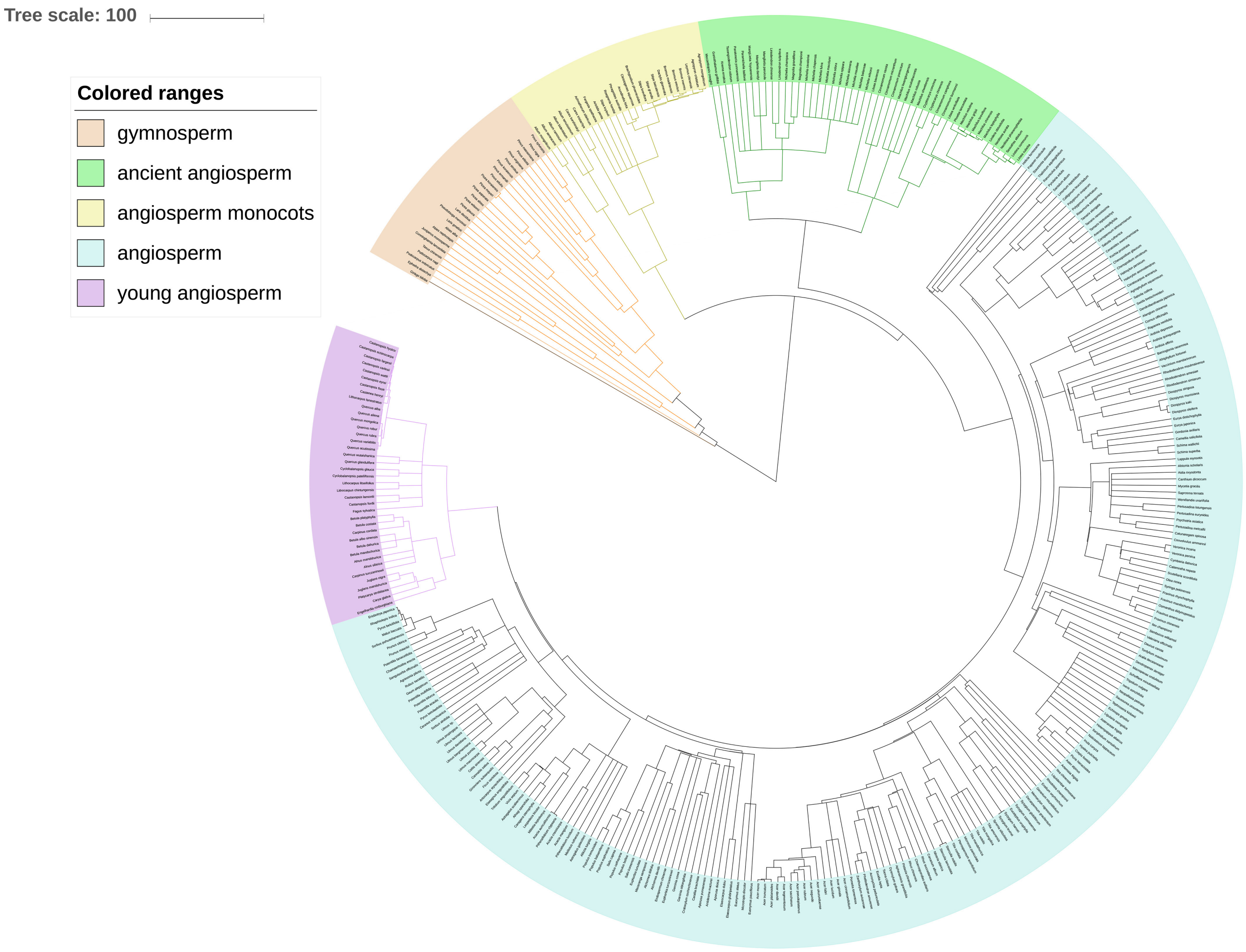
c

Desert	Desert	Grassland	Mediterranean	Boreal	Temperate	Subtropical
Grassland	0.07					
Mediterranean	0.16	<0.05*				
Boreal	0.25	<0.05*	1.00			
Temperate	0.20	0.10	<0.05*	<0.05*		
Subtropical	<0.001***	<0.05*	<0.001**	<0.001**	<0.001**	
Tropical	<0.001**	<0.05*	<0.001**	<0.01**	<0.001**	0.45

Extended Data Figure 5 | Distribution of first-order-root diameter for woody and herbaceous plants across biomes. a, Woody plant root diameter deceases from tropical to desert biomes, with the most frequent occurrence of coarse-root ancestral woody species in tropical and subtropical biomes. **b**, Herbaceous plant root diameters do not display a clear trend across biomes. In both panels, the letters ‘a’, ‘b’ and ‘c’ denote significant differences ($P < 0.05$) between biomes based on a linear mixed effects model (generated using the lmer function in R) with species included as a random effect. Diameter was first $\log_{10}$-transformed to correct for non-normality. Each point represents a species-specific observation at one site. The background violin plot characterizes the distribution of points in each biome. **c**, Pairwise comparisons for equal variance in first-order-root diameter using Levene’s test. Levene’s test is used for testing the homogeneity of variance, and is used here to explain biome differences in variance of root diameter.



Extended Data Figure 6 | Frequency distributions of nine root functional traits. a–c, Cyan bars identify the distribution of herbaceous plants, yellow bars identify woody plants, and green colour is where two distributions overlap. *n*, total number of species; *s*, skewness of all data.



Extended Data Figure 7 | Phylogenetic tree of 365 taxa in the study. The oldest taxonomic groups are highlighted in orange (gymnosperms), yellow (monocotyledons) and green (for example, Magnoliales, Lauraceae). The youngest taxonomic groups are highlighted in purple (for example, Betulaceae, Fagaceae).

Extended Data Table 1 | Summary of ten functional traits of first order roots

Functional traits	Abbreviation	*n*	Mean	SE	Median	Min	Max	CV%*	Blomberg's K†
Root diameter (mm)	Diam	369	0.29	0.01	0.24	0.08	1.01	57%	0.11***
Specific root length (m g^{-1})	SRL	323	110.9	6.6	72.2	4.4	657.1	107%	0.10***
Root length (mm)	Length	270	8.10	0.31	7.63	0.70	25.7	64%	0.08**
Root tissue density (g cm^{-3})	RTD	315	0.32	0.01	0.29	0.06	0.95	52%	0.03
Root N concentration (mg g^{-1})	RootN	284	19.3	0.3	19.0	5.6	50.3	30%	0.03
Root C concentration (mg g^{-1})	RootC	252	473.0	3.7	467.0	307.7	684.4	12%	0.01
Root C:N ratios	RootCN	272	25.0	0.7	23.7	2.1	77.8	48%	0.03
Mycorrhizal colonization (%)	MC	137	52	2	55	0	100	54%	0.32**
Root median lifespan (days)‡	RL	75	272	24	234	34	924	77%	0.04
Root N uptake rates (μg N g^{-1} Root h^{-1})§	RNU	141	52.5	7.3	18	0	511	164%	0.02

n, number of species analysed.

*CV%, the coefficient of variance.

†Blomberg's *K* value; the strength of phylogenetic signal with significance level (**$P < 0.01$; ***$P < 0.001$).

‡Root median lifespans were derived from direct field observations using minirhizotrons and root windows (see Methods).

§Root nitrogen uptake rates were compiled from 43 publications[59–101] and unpublished data.

Extended Data Table 2 | Principal component analyses of global root functional trait data

	Main analysis		Excl.MC		Gapped.MC.data		Gap-filled	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Eigenvalue	3.21	1.37	2.25	1.32	2.71	1.33	2.35	1.57
Proportion of Variance	46%	19%	38%	22%	39%	19%	34%	22%
Root diameter (Diam)	0.94	0.09	0.78	-0.48	0.84	-0.37	0.88	-0.25
Root length (Length)	0.71	-0.19	0.82	-0.12	0.72	-0.03	0.13	-0.24
Specific root length (SRL)	-0.87	0.05	-0.81	-0.09	-0.81	-0.21	-0.76	-0.29
Root tissue density (RTD)	-0.64	-0.5	-0.09	0.84	-0.15	0.85	-0.09	0.88
Root C concentration (RootC)	-0.2	0.60	0.39	0.12	0.07	0.01	0.07	0.07
Root N concentration (RootN)	0.09	0.83	-0.11	-0.72	-0.02	-0.71	-0.13	-0.65
Mycorrhizal colonization (MC)	0.79	-0.07			0.79	-0.14	0.87	0.06

Eigenvalues and loading scores of principal components 1 and 2 (PC1 and PC2) in four different principal component analyses. For all four analyses, the loading scores show that the PC1 axis is composed of root diameter, SRL and length, and the PC2 axis is mainly influenced by root nitrogen concentration. The 'main analysis' is reported in Extended Data Fig. 1, and was conducted on 104 species for which all 7 traits were available. The 'Excl.MC' analysis excluded mycorrhizal colonization, which enabled us to increase the number of species to 217. In the 'Gapped.MC.data' analysis, we filled gaps in the mycorrhizal colonization data based on the relationships in Fig. 1c, resulting in 217 species being analysed. Finally, in the 'Gap-filled' analysis, we filled all gaps in trait values based either on the relationships in Fig. 1a, c or on interpolations using the MICE multiple imputation method. The results of all four principal component analyses show the same overall result, which indicates that the pattern shown in Extended Data Fig. 1 is robust.

Extended Data Table 3 | Spearman's correlation coefficients and phylogenetically independent contrasts among eight root functional traits

	Diam	SRL	RTD	Length	RootN	RootC	RootCN	MC
Root diameter (Diam)		-0.87***	-0.26***	0.34***	0.21***	0.04	-0.05	0.67***
Specific root length (SRL)	-0.59***		-0.09	-0.25***	-0.22***	0.15**	-0.07	-0.69***
Root tissue density (RTD)	-0.27***	-0.15**		-0.33***	-0.05	-0.30***	0.13*	0.06
Root length (Length)	0.32***	-0.15**	-0.29***		0.37***	0.04	0.05	0.06
Root C concentration (RootC)	0.19***	-0.12*	-0.09	0.32***		0.14**	0.23***	0.14**
Root N concentration (RootN)	0.09	0.11*	-0.24***	0.03	0.21***		-0.65***	-0.08
Root C:N ratios (RootCN)	0.05	-0.08	0.05	0.08	0.16**	-0.51***		-0.02
Mycorrhizal colonization (MC)	0.32***	-0.25***	0.01	-0.03	-0.06	-0.18***	0.06	

Spearman's correlation coefficients (upper diagonal) and phylogenetically independent contrasts (lower diagonal) among eight root functional traits for 365 species. Significant correlations are indicated; $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$.

doi:10.1038/nature25740

Circadian clock neurons constantly monitor environmental temperature to set sleep timing

Swathi Yadlapalli[1]*, Chang Jiang[2]*, Andrew Bahle[1]†, Pramod Reddy[2], Edgar Meyhofer[2] & Orie T. Shafer[1]

Circadian clocks coordinate behaviour, physiology and metabolism with Earth's diurnal cycle[1,2]. These clocks entrain to both light and temperature cycles[3], and daily environmental temperature oscillations probably contribute to human sleep patterns[4]. However, the neural mechanisms through which circadian clocks monitor environmental temperature and modulate behaviour remain poorly understood. Here we elucidate how the circadian clock neuron network of *Drosophila melanogaster* processes changes in environmental temperature. *In vivo* calcium-imaging techniques demonstrate that the posterior dorsal neurons 1 ($\mathrm{DN1_p}$s), which are a discrete subset of sleep-promoting clock neurons[5], constantly monitor modest changes in environmental temperature. We find that these neurons are acutely inhibited by heating and excited by cooling; this is an unexpected result when considering the strong correlation between temperature and light, and the fact that light excites clock neurons[6]. We demonstrate that the $\mathrm{DN1_p}$s rely on peripheral thermoreceptors located in the chordotonal organs[7,8] and the aristae[9]. We also show that the $\mathrm{DN1_p}$s and their thermosensory inputs are required for the normal timing of sleep in the presence of naturalistic temperature cycles. These results identify the $\mathrm{DN1_p}$s as a major gateway for temperature sensation into the circadian neural network, which continuously integrates temperature changes to coordinate the timing of sleep and activity.

Circadian clocks display a remarkable combination of imperturbability and sensitivity with regard to temperature. Circadian clocks maintain free-running periods that are strongly temperature-compensated[10], and brief temperature pulses do not reset the phase of free-running circadian rhythms[11]. Nevertheless, circadian rhythms entrain to low-amplitude step-function temperature cycles[12,13]. The neural mechanisms by which temperature modulates circadian rhythms are not well understood. Here we use *Drosophila melanogaster*, an organism with a highly conserved yet relatively simple circadian clock neuron network[14], to investigate how the clock neuron network processes changes in environmental temperature.

To determine whether and how the various clock neuron classes (Fig. 1a) respond to changes in temperature, we performed experiments using CaMPARI[15], a fluorescent protein in which photoconversion from green to red is proportional to $\mathrm{Ca^{2+}}$ levels. We affixed individual flies to a Peltier module (Fig. 1b) and exposed them to heating, cooling or constant temperature in the presence of 405-nm light (Extended Data Fig. 1). Upon heating, the $\mathrm{DN1_p}$s and DN2s displayed reduced photoconversion compared to controls, whereas the remaining classes showed no significant changes (Fig. 1c, d). Cooling resulted in increased photoconversion within the $\mathrm{DN1_p}$, DN2 and DN3 classes, whereas the remaining classes showed no changes (Fig. 1c, e). These results suggest—in contrast to longstanding expectations[10,11]—that nodes within the circadian network are unexpectedly sensitive to brief changes in temperature, and reveal that $\mathrm{DN1_p}$s are acutely inhibited by heating and excited by cooling. Levels of environmental light and temperature rise and fall concurrently, with temperature lagging slightly behind light. Our results show that light and temperature are processed in distinct ways in the clock neuron network: light is excitatory[6] and heating is inhibitory. Cooling had no effect on clock neurons in isolated brains (Extended Data Fig. 2), suggesting a requirement for peripheral thermosensors.

Next, we examined the kinetics of temperature response by monitoring intracellular $\mathrm{Ca^{2+}}$ concentrations ($[\mathrm{Ca^{2+}}]_i$) using the $\mathrm{Ca^{2+}}$ sensor

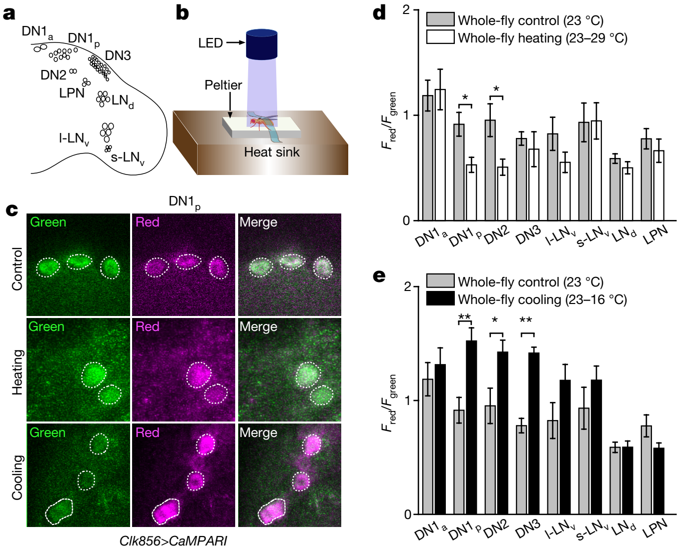


Figure 1 | *Drosophila* circadian clock neurons respond to temperature changes. a, Schematic of the *Drosophila* circadian clock network, with the major classes of pacemaker labelled. **b**, Schematic of the setup for the CaMPARI experiments. **c**, Representative images of CaMPARI expression (driven by the pan-clock neuron driver *Clk856-GAL4*) in the $\mathrm{DN1_p}$s in response to constant temperature (23 °C, top), heating (23–29 °C, middle) and cooling (23–16 °C, bottom). Results are representative of ten independent experiments (brains) for each treatment. **d**, **e**, Quantification of the ratio of red to green CaMPARI fluorescence in various clock neuron classes when adult flies were heated (**d**) or cooled (**e**). The grey bars in **d** and **e** denote basal photoconversion at constant temperature (12 $\mathrm{DN1_a}$s, 77 $\mathrm{DN1_p}$s, 13 DN2s, 26 DN3s, 29 l-$\mathrm{LN_v}$s, 7 s-$\mathrm{LN_v}$s, 47 $\mathrm{LN_d}$s, 19 LPNs); white bars in **d** denote heating responses (10 $\mathrm{DN1_a}$s, 89 $\mathrm{DN1_p}$s, 20 DN2s, 14 DN3s, 26 l-$\mathrm{LN_v}$s, 12 s-$\mathrm{LN_v}$s, 43 $\mathrm{LN_d}$s, 19 LPNs); black bars in **e** denote cooling responses (10 $\mathrm{DN1_a}$s, 73 $\mathrm{DN1_p}$s, 16 DN2s, 11 DN3s, 32 l-$\mathrm{LN_v}$s, 11 s-$\mathrm{LN_v}$s, 45 $\mathrm{LN_d}$s, 20 LPNs). In all bar plots, histograms represent mean ± s.e.m; 10–13 brains were analysed in each experiment. $^{**}P < 0.005$; $^{*}P < 0.05$; unpaired two-tailed Student's *t*-test. Specific *P* values are reported in the Source Data for this figure. $\mathrm{DN1_a}$s, anterior dorsal neurons 1; l-$\mathrm{LN_v}$s, large ventral lateral clock neurons; s-$\mathrm{LN_v}$s, small ventral lateral clock neurons; $\mathrm{LN_d}$s, dorsal lateral neurons; LPNs, lateral posterior neurons.

[1] Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. [2] Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA. †Present address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

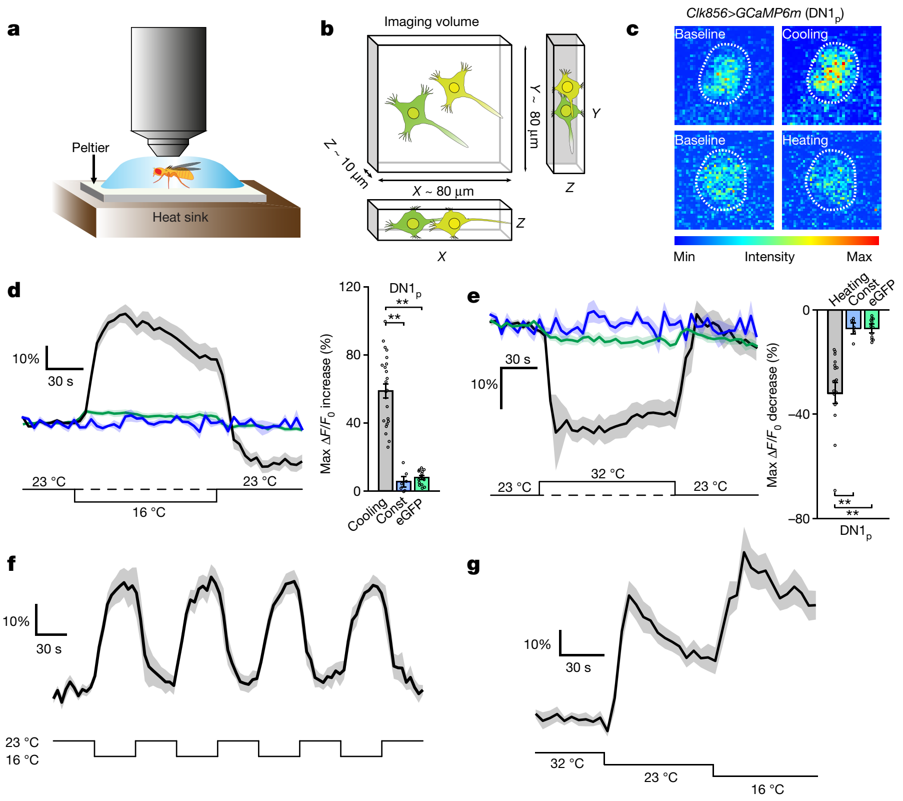


Figure 2 | DN1$_p$s are acutely excited by cooling and inhibited by heating. **a**, Schematic of the setup for *in vivo* calcium-imaging experiments. **b**, Schematic of the 3D imaging of neurons (see Methods). **c**, Representative images of the fluorescence response of GCaMP6m in DN1$_p$s in response to cooling (23–16 °C, top right), heating (23–32 °C, bottom right) and constant temperature (23 °C, left), based on 7 biologically independent samples. GCaMP is driven by the pan-clock neuron driver *Clk856-GAL4*. **d**, Averaged responses of GCaMP6m in DN1$_p$s to cooling (black, $n = 22$) and constant temperature (blue, $n = 7$); response of enhanced green fluorescent protein (eGFP) to cooling (green, $n = 14$). Data are presented as average maximum increase of $\Delta F/F_0$ (%) ± s.e.m. in the histogram; one-way analysis of variance (ANOVA), $F_{2,42} = 60.36$, $P = 8.32 \times 10^{-13}$. **e**, Averaged responses of GCaMP6m in DN1$_p$s to heating (black, $n = 18$) and constant temperature (blue, as in **d**); response of eGFP to heating (green, $n = 12$). Data are presented as average maximum decrease of $\Delta F/F_0$ (%) ± s.e.m. in the histogram; one-way ANOVA, $F_{2,39} = 18.27$, $P = 3.03 \times 10^{-6}$. **f**, DN1$_p$s respond reliably to repeated cold and hot stimuli ($n = 9$). **g**, DN1$_p$s physiologically track two successive cooling steps (32–23–16 °C) ($n = 10$). $^{**}P < 0.005$; unpaired two-tailed Student's *t*-test. For all GCaMP6m plots, $\Delta F/F_0$ (%) indicates the percentage of fluorescence change compared with baseline, dark lines indicate average ΔF, shaded areas indicate s.e.m. The temperature profile is shown at the bottom of the GCaMP6m plot. n = number of neurons, with individual data points shown if $n < 10$; 5–15 brains were analysed in each experiment. In **d** and **e**, the ANOVA was performed across the entire group, and Tukey's honest significant difference tests were conducted in a pairwise fashion with P values indicated on the graphs. $^{**}P < 0.005$. Specific P values are reported in the Source Data for this figure.

GCaMP6m (ref. 16) (Fig. 2a, 2b, Extended Data Fig. 3). In agreement with the results from the CaMPARI experiments, the DN1$_p$s showed increases in $[Ca^{2+}]_i$ during cooling and decreases in $[Ca^{2+}]_i$ during heating (Fig. 2c–e), whereas the LN$_d$ neurons displayed no significant responses (Extended Data Fig. 3c, d). The DN1$_p$s exhibited significant responses even to modest (4 °C) temperature changes (Extended Data Fig. 3e) and responded repeatedly to rapid cycles of cooling and heating (Fig. 2f). To understand whether the DN1$_p$s were tracking the environmental temperature, or were simply excited when the temperature decreased below a threshold, we measured the responses of the DN1$_p$s to a double cooling step, and found that they responded robustly to both steps (Fig. 2g). DN1$_p$ cooling responses in flies that were entrained to light:dark cycles were higher in the morning than in the evening in a manner dependent on the circadian clock (Extended Data Fig. 3f), consistent with previous findings that DN1$_p$s are more excitable in the morning[17]. These results establish that the DN1$_p$s constantly monitor environmental temperature, are excited by cooling and inhibited by heating. At first, these results may appear to contradict recent work describing increases in Ca^{2+} levels within clock cells in response to heating[5]. However, these studies did not measure the acute physiological responses of clock neurons to brief temperature changes; rather they examined the effects of prolonged increased temperature. Our findings reveal that the circadian network transduces brief and transient temperature changes and prolonged increases in temperature in distinct ways.

In *Drosophila,* thermoreceptors reside in structures in the antennae, called the aristae[9,18] and the sacculi[19], and within chordotonal organs in the body[7,8] (Fig. 3a). Each arista contains both cold-sensitive and heat-sensitive cells[9], whereas a sacculus contains only cold-sensitive cells[19]. Each chordotonal organ contains heat-activated cells, which are required for behavioural synchronization to low-amplitude, step-function temperature cycles[8]. In addition, a group of *Transient receptor potential cation channel A1* (*TrpA1*)-expressing anterior cells in the brain act as internal heat sensors[20]. We performed GCaMP imaging experiments on flies from which the aristae or the body had been removed. The responses to cooling and heating were attenuated when the aristae or the body were removed, with only the ipsilateral arista required for normal-amplitude cooling responses (Fig. 3b–d). In isolated brains, which lack both the aristae and the body, DN1$_p$ cooling responses were completely abolished, whereas slight heat-induced decreases in GCaMP fluorescence persisted (Fig. 3b, c, Extended Data Fig. 4a, b). These persistent heating responses appear to be an artefact, as they persist in the absence of neuronal firing, synaptic transmission and anterior-cell neuron function (Extended Data Fig. 4c).

Next, using GCaMP imaging, we measured the responses of DN1$_p$s in *nocte^1* mutants, which have defects in both chordotonal organ

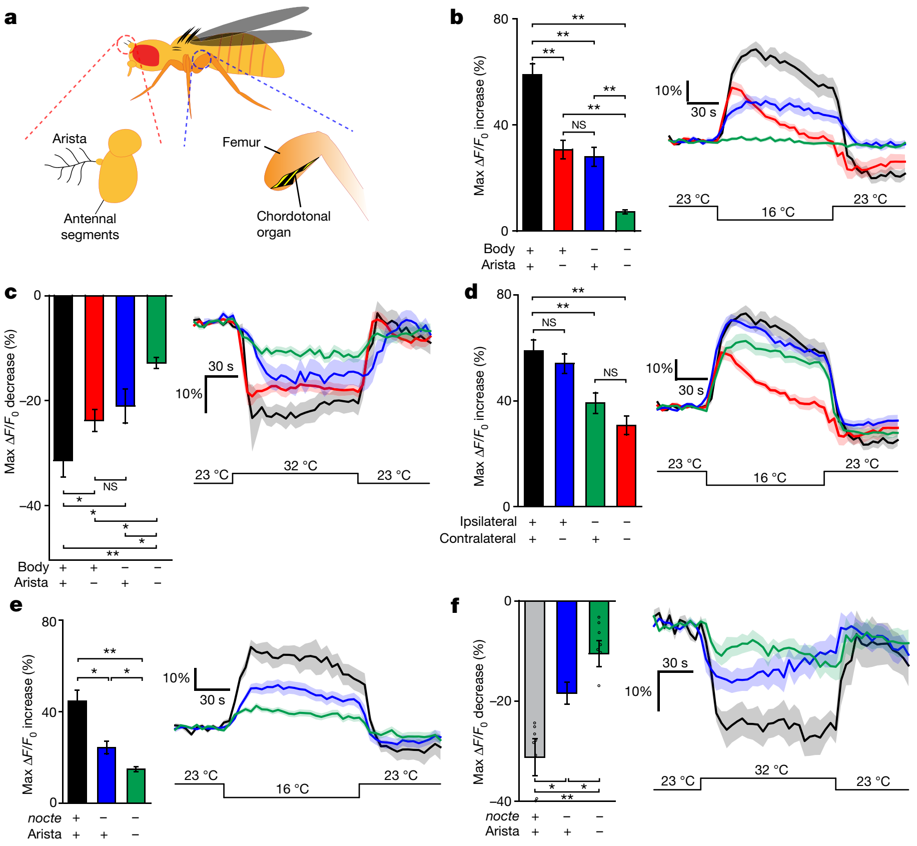


Figure 3 | Thermoreceptors in the aristae and the chordotonal organs are required for the responses of DN1$_{p}$s to temperature changes. **a**, Schematic of the anatomy of temperature sensation in *Drosophila*. **b**, Averaged responses of GCaMP6m in DN1$_{p}$s to cooling in the intact fly (black, $n=22$), after removal of aristae (red, $n=11$), after removal of the body (blue, $n=10$) and in isolated brains (green, $n=32$). Data are presented as average maximum increase of $\Delta F/F_0$ (%) ± s.e.m. in the histograms; one-way ANOVA, $F_{3,74}=67.60$, $P=1.12\times10^{-13}$. **c**, Averaged responses of GCaMP6m in DN1$_{p}$s to heating in the intact fly (black, $n=18$), after removal of aristae (red, $n=16$), after removal of the body (blue, $n=12$) and in isolated brains (green, $n=25$). Data are presented as average maximum decrease of $\Delta F/F_0$ (%) ± s.e.m. in the histograms; one-way ANOVA, $F_{3,70}=14.71$, $P=1.85\times10^{-7}$. **d**, Averaged responses of GCaMP6m in DN1$_{p}$s to cooling, after the removal of one or both the aristae: intact flies (black, $n=22$), contralateral arista removed (blue, $n=21$), ipsilateral arista removed (green, $n=23$) and both aristae removed (red, $n=11$). Histogram quantification as in **b**; one-way ANOVA, $F_{3,76}=8.80$, $P=4.7\times10^{-5}$. **e**, Averaged responses of GCaMP6m in DN1$_{p}$s to cooling in *nocte*$^{+}$ (black, $n=10$), *nocte*^{1} mutants (blue, $n=22$), and *nocte*^{1} mutants from which arista were removed (green, $n=21$). Histogram quantification as in **b**; one-way ANOVA, $F_{2,52}=23.89$, $P=5.22\times10^{-8}$. **f**, Averaged responses of GCaMP6m in DN1$_{p}$s to heating in *nocte*$^{+}$ (grey, $n=8$), *nocte*^{1} mutants (blue, $n=12$), and *nocte*^{1} mutants from which aristae were removed (green, $n=8$). Histogram quantification as in **c**; one-way ANOVA, $F_{2,27}=12.31$, $P=1.9\times10^{-4}$. In all of the panels, the GCaMP6m plots shown on the right indicate the percentage of fluorescence change compared with the baseline ($\Delta F/F_0$ (%)); dark lines indicate average ΔF, shaded areas indicate s.e.m. The temperature profile for each experiment is shown at the bottom of the GCaMP6m plot. n = number of neurons, 5–10 brains were analysed in each experiment. Tukey's honest significant difference tests were conducted in pairwise fashion. $*P<0.05$, $**P<0.005$; NS, not significant. Specific P values are reported in the Source Data for this figure.

structure and temperature entrainment[7,8]. The responses of DN1$_{p}$s to temperature changes are reduced in *nocte*^{1} mutants, and almost completely abolished in *nocte*^{1} mutants from which the aristae were removed (Fig. 3e, f). Finally, RNAi-mediated knockdown of *nocte* expression in the chordotonal organs caused a *nocte*^{1}-like reduction in the responses of DN1$_{p}$s to temperature (Extended Data Fig. 4d, e). The chordotonal organs therefore appear to be the primary source of temperature input from the body to the DN1$_{p}$s. We conclude that peripheral thermoreceptors in both the aristae and the chordotonal organs contribute to the thermal sensitivity of DN1$_{p}$s.

We turned our attention to the role of the DN1$_{p}$s in the timing of sleep and activity under constantly cycling temperature, as such cycles are more similar to those found in natural environments[4,21] than the step-function temperature cycles typically used in the laboratory. *Drosophila* locomotor activity readily entrains to gradually ramping temperature cycles between 18 °C and 25 °C that mimic rising temperatures during the morning and falling temperatures during the evening[22] (Fig. 4a, b). Under these conditions, flies display locomotor activity that slowly increases throughout the heating phase followed by a precipitous increase in sleep that anticipates the onset of cooling[22] (Fig. 4a, f, Extended Data Figs 5a–d, 6). This sharp transition to sleep, which persists under free-running conditions (Extended Data Fig. 7), is notable considering that the maximal rate of temperature change was only around 0.6 °C h^{-1}. Flies entrained to ramping temperature cycles respond to unexpected bouts of heating and cooling with increased activity and sleep, respectively (Extended Data Fig. 5e–g). In the absence of clock function, flies are sensitive to the heating and cooling transitions of ramping temperature cycles: *per*^{01} mutants, for example, precipitously increase their activity at the onset of heating and decrease their activity at the onset of cooling, failing to gradually increase activity throughout the heating phase or to anticipate cooling (Fig. 4c, f, Extended Data Fig. 6). Outside the laboratory, flies display a mid-afternoon activity peak, with a relationship to the environmental temperature cycle that is markedly similar to that seen for ramping temperature cycles[21,22].

We found that the *nocte*^{1} mutants and flies that lacked aristae displayed reduced anticipation to cooling and a slow transition to sleep, and that *nocte*^{1} mutant flies from which aristae were removed showed clearly

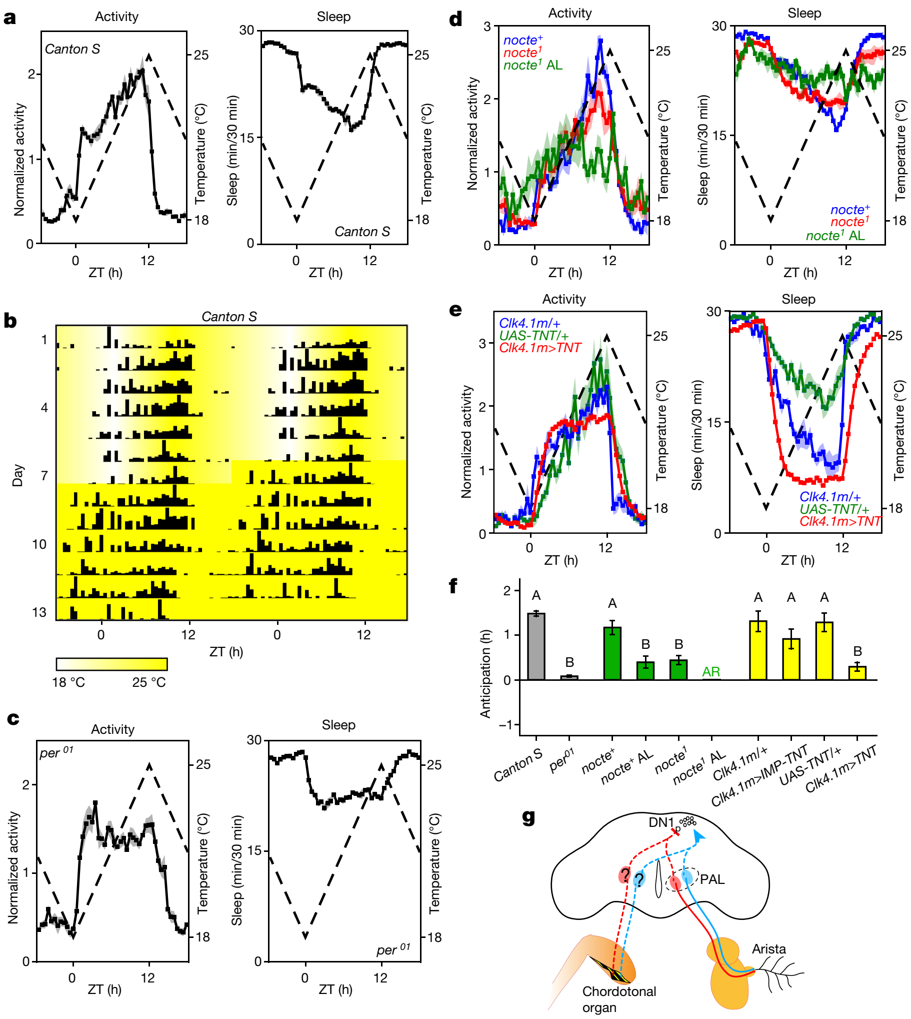


Figure 4 | DN1$_p$s are required for the synchronization of behaviour to ramping temperature cycles. For all behavioural experiments, flies were entrained to temperature cycles (see Methods). **a**, **c**–**e**, Averaged population activity plots for the last four days of temperature cycles (left), and averaged sleep plots (right) for the following genotypes: **a**, *Canton S* ($n = 214$), **c**, *per^{01}* ($n = 96$), **d**, *nocte^{+}* (blue, $n = 46$), *nocte^{1}* (red, $n = 56$), aristae-less (AL) *nocte^{1}* (green, $n = 44$), **e**, *Clk4.1m>UAS-TNT* (red, $n = 100$), *Clk4.1m/+* (dark blue, $n = 17$), *UAS-TNT/+* (green, $n = 23$). Dashed lines represent temperature. Dark lines indicate average activity, shaded areas indicate s.e.m., n = number of flies. ZT, Zeitgeber time. **b**, Representative double-plotted actogram of an individual wild-type *Canton S* fly. The temperature gradient is shown at the bottom of the actogram. **f**, Cooling anticipation of the different genotypes, generated using the sleep data from all the flies from experiments represented in **a**, **c**–**e**. n values are reported in **a**, **c**–**e**. For all of the histograms apart from that of *nocte^{1}* AL, data are presented as population average of anticipation ± s.e.m. For *nocte^{1}* AL flies, AR indicates that the majority of the flies are arrhythmic, making the calculation of anticipation indices impossible (see Methods). One-way ANOVA: $F_{1,309} = 230.3$, $P < 0.0001$ among grey bars; $F_{2,129} = 19.97$, $P = 3.99 \times 10^{-5}$ among green bars; $F_{3,171} = 15.16$, $P = 2.23 \times 10^{-8}$ among yellow bars. The letters A, B and C above the histograms denote significantly different means within each of the experimental groups, Tukey's honest significant difference tests were conducted in a pairwise fashion, $P < 0.005$ for all. Specific P values are reported in the Source Data for this figure. **g**, A model of the temperature input pathways into the *Drosophila* circadian clock network. PAL, posterior antennal lobe.

abnormal behaviour under ramping temperature cycles (Fig. 4d, f, Extended Data Figs 6, 8). Although the *nocte^{1}* flies that lacked aristae were able to entrain to light:dark cycles, a majority of them were not able to entrain to temperature cycles and were arrhythmic under constant conditions after the temperature-entrainment period (Extended Data Fig. 8). Similar defects were observed in flies that both lacked aristae and in which *nocte* had been knocked down in the chordotonal organs (Extended Data Fig. 8). Therefore, the thermoreceptors that are required for the temperature sensitivity of DN1$_p$s are also required for the normal entrainment to ramping temperature cycles, and flies that lack both aristae and chordotonal function are 'circadian blind'[23] to such cycles.

To determine the extent to which the DN1$_p$s are required for the normal pattern of sleep and activity in the presence of ramping temperature cycles, we expressed tetanus-toxin (TNT) to block synaptic transmission[24] in approximately half of the DN1$_p$s using the *Clk4.1m-GAL4* driver[25]. The effects were markedly reminiscent of the effects of the *per^{01}* mutation (Fig. 4c, e, f, Extended Data Fig. 6). Consistent with previous reports that identified DN1$_p$s as sleep-promoting neurons[5], *Clk4.1M/UAS-TNT* flies were found to sleep less than control flies during the heating phase, but surprisingly, they displayed normal levels of sleep during the cooling phase (Extended Data Fig. 7). We also tested *glass^{60j}* mutants, which lack all DN1$_p$s in addition to all external photoreceptors[26], and found that they neither gradually increase their activity during heating nor display cooling-associated increases in sleep (Extended Data Fig. 9), despite the presence of aristae and chordotonal organs. The expression of the toxin within DN2s did not affect the rhythmic behaviour of flies under ramping temperature cycles (Extended Data Figs 6, 9). Taken together, the results from our behavioural experiments are consistent with those of previous work[8], which established a role for DN1$_p$s in entrainment to step-function temperature cycles, and further revealed that DN1$_p$s have a critical role in the timing of sleep and activity under constantly changing temperature cycles. We note, however, that these neurons are not necessary for the response to unexpected bouts of heating and cooling (Extended Data Fig. 9d, e). Finally, the activity peak under ramping temperature cycles is controlled by the $DN1_p$ clocks: it is advanced or delayed when the circadian clock is sped up or slowed down specifically in these neurons, respectively (Extended Data Fig. 10).

Our results reveal a central role for $DN1_p$s in temperature entrainment and the generation of sleep and activity rhythms in the presence of constantly changing environmental temperature cycles, through the integration of temperature inputs from two distinct peripheral thermosensors (Fig. 4g). Our finding that $DN1_p$s are inhibited by heating and excited by cooling provides important insights into how the clock-neuron network of the fly is likely to coordinate behaviour under constantly changing environmental temperatures. We propose that the slow and gradual increase in activity observed during the heating phase of the temperature cycle is promoted through increasing inhibition of the $DN1_p$s, neurons recently shown to promote sleep[5], whereas the precipitous increase in sleep preceding the onset of cooling is mediated by the excitation of $DN1_p$s. Our work reveals how a circadian clock modulates sleep and activity by constantly integrating temperature signals into its neural network. The duration and timing of sleep are closely correlated to both environmental temperature cycles[4] and clock-controlled body temperature rhythms[27,28] throughout the animal kingdom. We therefore expect that the mechanisms reported here will be broadly relevant to sleep, including sleep in humans.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 December 2016; accepted 10 January 2018.
Published online 21 February 2018.

1. Aschoff, J. *Biological Rhythms* (Plenum, 1981).
2. Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96,** 271–290 (1999).
3. Golombek, D. A. & Rosenstein, R. E. Physiology of circadian entrainment. *Physiol. Rev.* **90,** 1063–1102 (2010).
4. Yetish, G. *et al.* Natural sleep and its seasonal variations in three pre-industrial societies. *Curr. Biol.* **25,** 2862–2868 (2015).
5. Guo, F. *et al.* Circadian neuron feedback controls the *Drosophila* sleep–activity profile. *Nature* **536,** 292–297 (2016).
6. Fogle, K. J., Parson, K. G., Dahm, N. A. & Holmes, T. C. CRYPTOCHROME is a blue-light sensor that regulates neuronal firing rate. *Science* **331,** 1409–1413 (2011).
7. Sehadova, H. *et al.* Temperature entrainment of *Drosophila*'s circadian clock involves the gene *nocte* and signaling from peripheral sensory tissues to the brain. *Neuron* **64,** 251–266 (2009).
8. Chen, C. *et al. Drosophila* Ionotropic Receptor 25a mediates circadian clock resetting by temperature. *Nature* **527,** 516–520 (2015).
9. Gallio, M., Ofstad, T. A., Macpherson, L. J., Wang, J. W. & Zuker, C. S. The coding of temperature in the *Drosophila* brain. *Cell* **144,** 614–624 (2011).
10. Pittendrigh, C. S. On temperature independence in the clock system controlling emergence time in *Drosophila*. *Proc. Natl Acad. Sci. USA* **40,** 1018–1029 (1954).
11. Edery, I., Rutila, J. E. & Rosbash, M. Phase shifting of the circadian clock by induction of the *Drosophila* period protein. *Science* **263,** 237–240 (1994).
12. Wheeler, D. A., Hamblen-Coyle, M. J., Dushay, M. S. & Hall, J. C. Behavior in light-dark cycles of *Drosophila* mutants that are arrhythmic, blind, or both. *J. Biol. Rhythms* **8,** 67–94 (1993).
13. Buhr, E. D., Yoo, S. H. & Takahashi, J. S. Temperature as a universal resetting cue for mammalian circadian oscillators. *Science* **330,** 379–385 (2010).
14. Nitabach, M. N. & Taghert, P. H. Organization of the *Drosophila* circadian control circuit. *Curr. Biol.* **18,** R84–R93 (2008).
15. Fosque, B. F. *et al.* Neural circuits. Labeling of active neural circuits *in vivo* with designed calcium integrators. *Science* **347,** 755–760 (2015).
16. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
17. Flourakis, M. *et al.* A conserved bicycle model for circadian clock control of membrane excitability. *Cell* **162,** 836–848 (2015).
18. Liu, W. W., Mazor, O. & Wilson, R. I. Thermosensory processing in the *Drosophila* brain. *Nature* **519,** 353–357 (2015).
19. Barbagallo, B. & Garrity, P. A. Temperature sensation in *Drosophila*. *Curr. Opin. Neurobiol.* **34,** 8–13 (2015).
20. Hamada, F. N. *et al.* An internal thermal sensor controlling temperature preference in *Drosophila*. *Nature* **454,** 217–220 (2008).
21. Vanin, S. *et al.* Unexpected features of *Drosophila* circadian behavioural rhythms under natural conditions. *Nature* **484,** 371–375 (2012).
22. Currie, J., Goda, T. & Wijnen, H. Selective entrainment of the *Drosophila* circadian clock to daily gradients in environmental temperature. *BMC Biol.* **7,** 49–67 (2009).
23. Helfrich-Förster, C., Winter, C., Hofbauer, A., Hall, J. C. & Stanewsky, R. The circadian clock of fruit flies is blind after elimination of all known photoreceptors. *Neuron* **30,** 249–261 (2001).
24. Sweeney, S. T., Broadie, K., Keane, J., Niemann, H. & O'Kane, C. J. Targeted expression of tetanus toxin light chain in *Drosophila* specifically eliminates synaptic transmission and causes behavioral defects. *Neuron* **14,** 341–351 (1995).
25. Zhang, Y., Liu, Y., Bilodeau-Wentworth, D., Hardin, P. E. & Emery, P. Light and temperature control the contribution of specific DN1 neurons to *Drosophila* circadian behavior. *Curr. Biol.* **20,** 600–605 (2010).
26. Klarsfeld, A. *et al.* Novel features of cryptochrome-mediated photoreception in the brain circadian clock of *Drosophila*. *J. Neurosci.* **24,** 1468–1477 (2004).
27. Kaneko, H. *et al.* Circadian rhythm of temperature preference and its neural control in *Drosophila*. *Curr. Biol.* **22,** 1851–1857 (2012).
28. Kryger, M. H., Roth, T. & Dement, W. C. *Principles and Practice of Sleep Medicine* 6th edn (Elsevier, 2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by a Damon Runyon Cancer Foundation postdoctoral fellowship (DR2231-15) to S.Y., a National Science Foundation (NSF) CBET grant (1509691) to P.R. and E.M., a University of Michigan M-cubed grant to E.M., P.R. and O.T.S., and a National Institutes of Health NINDS grant (R01NS077933) and an NSF IOS grant (1354046) to O.T.S. We thank N. Glossop, P. Hardin, M. Rosbash and R. Stanewsky for fly stocks. We also thank M. Rosbash, L. Griffith and P. Garrity for discussions about our work and M. de la Paz Fernández for reading the manuscript.

Author Contributions The work was conceived by S.Y. and O.T.S. The calcium-imaging and behavioural experiments were performed by S.Y. and C.J., under the supervision of O.T.S. C.J. and S.Y. designed and built the Peltier and CaMPARI setups under the guidance of P.R. and E.M. S.Y., C.J. and A.B. conducted CaMPARI experiments. Data analysis was performed by S.Y. and C.J. The manuscript was written by S.Y., C.J. and O.T.S. with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to O.T.S. (oshafer@umich.edu).

Reviewer Information *Nature* thanks P. Emery, R. Stanewsky and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. We followed well-established standards in this research field to conduct our *Drosophila* live imaging and circadian behaviour experiments. No statistical calculations were used to predetermine sample sizes. Our sample sizes are similar to those generally used in this field of research. Samples were not randomized for the experiments and blinding was not used.

Fly strains. Flies were raised on cornmeal-yeast-sucrose food (recipe four from Bloomington Drosophila Stock Center, https://bdsc.indiana.edu/information/recipes/bloomfood.html under a 12 h:12 h light:dark cycle at 25 °C and 60–70% humidity. The following flies used in the study were previously described or obtained from the Bloomington Drosophila Stock Center: *Canton S*, *Clk856-GAL4* (ref. 29) (expressed in all clock neuron classes), *UAS-CaMPARI* (ref. 15), *UAS-GCaMP6m* (ref. 16), *UAS-GFP*, *per*^{01} (ref. 30), *glass*^{60j} (ref. 26), *UAS-TRPA1* (ref. 20), *TRPA1* (refs 1,20), *UAS-TNT-E* (ref. 24) and *UAS-IMP-TNT-V1-B* (ref. 24) (inactive), *Clk4.1M-GAL4* (ref. 25) (expressed in subset of DN1$_{\mathrm{p}}$s), *Clk9M-GAL4;Pdf-gal80* (ref. 8) (expressed in DN2s), *nocte*^{1}: encodes truncated version of the NOCTE protein (obtained from R. Stanewsky, University Münster, Germany) (ref. 7), *nocte*$^{+}$, *F-GAL4* (ref. 7), *UAS-nocte-RNAi-1:3* (ref. 7), *LexAop-GCaMP3* (ref. 31) and *Cry-LexA* (ref. 32).

CaMPARI illumination setup. The CaMPARI setup consists of a Peltier element (430263-501, Laird Technologies) mounted under a 405 nm light-emitting diode (LED) device (LZ1-10UB00-00U7, LED Engin Inc.). The illumination system was constructed using a 60-mm cage system (Thorlabs Inc.) to ensure optical alignment. An LED with a nominal wavelength of 405 nm was mounted on a heat sink. The divergent beam emitted from the LED was collimated by a condenser lens. The beam diameter was reduced by a set of three lenses to uniformly illuminate an area with a diameter of approximately 6 mm on the Peltier element where the fly was immobilized with tape (Extended Data Fig. 1a). The intensity of the illuminating beam was controlled by a custom driver circuit applying constant current to the LED. At the base of the illumination system was a Peltier element, which was used to control the temperature of the fly, mounted on another heat sink.

Peltier setup. The temperature of the Peltier element was monitored by a thermistor and controlled with closed-loop feedback by a proportional–integral–derivative control system. The same heating and cooling stimuli were provided in both the CaMPARI and the GCaMP experiments. In the CaMPARI experiments, the fly was exposed to the ambient air. In these experiments, the actual fly temperature differed slightly from the Peltier element temperature owing to thermal resistance between the fly and the Peltier element (Extended Data Fig. 1c, d). The temperature of the fly was measured using a fine thermocouple (CHAL-005, Omega). This temperature change was achieved in around 7 s, owing to the intrinsic thermal time constant of the system composed of the Peltier element and the fly. In the GCaMP experiments, the fly was placed in saline, which is expected to reduce the thermal resistance substantially. Therefore, the temperature of the fly during heating and cooling was very close to the temperature of the Peltier element (Extended Data Fig. 1c, e). The presence of the saline solution increases the thermal mass and, therefore, the thermal time constant was increased to around 12 s in the GCaMP experiments. The isolated brain GCaMP and CaMPARI experiments were also performed in saline and underwent temperature changes similar to those observed in whole-fly GCaMP experiments.

CaMPARI protocol for whole-fly experiments. Transgenic flies with CaMPARI expressed throughout the circadian-clock network were generated by crossing *Clk856-GAL4* female to *UAS-CaMPARI* male flies. Ten- to fifteen-day-old female flies were used in all CaMPARI experiments, which were conducted between ZT6 and ZT10 on flies entrained to a 12 h:12 h light:dark cycle.

For whole-fly CaMPARI experiments, an individual fly was briefly anaesthetized on ice, and then immobilized on the Peltier element by applying adhesive tape across the thorax so that the abdomen and legs were in direct contact with the device (Extended Data Fig. 1a). Individual flies were exposed to hot stimulus (body temperature changed from 23 °C to 29 °C) or cold stimulus (body temperature changed from 23 °C to 16 °C) or constant temperature (23 °C), and 405-nm light simultaneously for 5 s (repeated for a total of 10 exposures) (Extended Data Fig. 1b). For heating experiments, the temperature of the fly was initially held at 23 °C for 30 s. The temperature was then raised to 29 °C. It took 7 s for the temperature of the fly to reach 29 °C. The temperature was held at 29 °C for 5 s while a 405-nm LED light was switched on. After the 5-s exposure to 405-nm light, the LED was turned off and the temperature was lowered back to 23 °C. This procedure was repeated for a total of 10 exposures. Cooling experiments were conducted with the same illumination protocol with temperatures lowered to 16 °C. Control experiments were performed at a constant temperature of 23 °C using the same illumination protocol to measure the basal photoconversion levels. All three groups of experiments (cooling, heating and control) were conducted on individual flies. Brains were dissected out immediately after the final LED stimulation and the clock neurons were imaged for green and red CaMPARI fluorescence. Brains were dissected in Ca^{2+}-free, haemolymph-like saline (HL3) consisting of (in mM): 70 NaCl, 5 KCl, 20 MgCl$_2$, 10 NaHCO$_3$, 5 D-(+)-trehalose, 115 sucrose, 5 HEPES, pH 7.1. Images of the neurons in red and green channels were acquired simultaneously using an Olympus FV1000 laser-scanning microscope (Olympus) using a ×60 objective (1.10 N/A W, FUMFL N, Olympus). Images were analysed in ImageJ and the intensity of the neurons was measured after background subtraction to obtain integrated intensities in the green (F_{green}) and red (F_{red}) channels respectively. The extent of CaMPARI photoconversion was determined as the ratio $F_{\mathrm{red}}/F_{\mathrm{green}}$.

CaMPARI protocol for isolated brain experiments. For isolated brain experiments, we found that control experiments on isolated brains (ten light pulses without any temperature changes) showed very high levels of basal photoconversion in all clock neuron classes (data not shown). Therefore, we modified the protocol for isolated brain CaMPARI experiments such that both of the following conditions were satisfied: control experiments using isolated wild-type brains show $F_{\mathrm{red}}/F_{\mathrm{green}}$ levels comparable to the levels observed when intact flies were exposed to ten light pulses (original CaMPARI protocol), and clock neurons ectopically expressing the temperature-sensitive *Drosophila* TRPA1 channel showed higher $F_{\mathrm{red}}/F_{\mathrm{green}}$ levels upon heating compared to the controls. To accomplish this, we exposed the brains to temperature changes and light pulses twice instead of ten times, as in the case of an intact fly (Extended Data Fig. 2).

GCaMP imaging. Live calcium imaging was conducted using 7–10-day-old *Clk856-GAL4>UAS-GCaMP6m* flies. The experiments were conducted between ZT6 and ZT10 on flies entrained to a 12 h:12 h light:dark cycle. We placed a 15 mm × 15 mm microscope cover glass on top of a Peltier device using adhesive tape. This cover glass served as a holder for saline solution. A hole of ~5 mm in diameter was drilled in the centre of the cover glass so that the fly could be placed in direct contact with the Peltier. Individual flies were immobilized on the Peltier element (as described in the section 'CaMPARI protocol for whole-fly experiments') and submerged in HL3 saline solution consisting of (in mM): 70 NaCl, 5 KCl, 20 MgCl$_2$, 1.5 CaCl$_2$ 10 NaHCO$_3$, 5 D-(+)-trehalose, 115 sucrose, 5 HEPES, pH 7.1. The head cuticle of the fly was removed to expose the dorsal brain. The fly was subjected to hot (body temperature changed from 23 °C to 32 °C) or cold (body temperature changed from 23 °C to 16 °C) stimuli while imaging fluorescing neurons using the laser scanning confocal microscope. To capture the movement of neurons during temperature changes, we sampled from a volume of brain (80 μm × 80 μm × 10 μm) using time-course Z-series imaging (15 optical sections for each time frame) (Extended Data Fig. 3a). To quantify the GCaMP fluorescence changes of clock-neuron cell bodies at a given time point, we selected a region of interest corresponding to the neuron of interest and measured the mean pixel intensity this region in all the optical sections (Extended Data Fig. 3a). The highest mean pixel intensity of the region of interest among all the optical sections and the intensities of the region of interest in the two adjacent sections (three red circles in Extended Data Fig. 3b) were chosen and a quadratic function was fitted to these three values. The maximum intensity value of the fitted quadratic function was used as the fluorescence intensity of the neuron at that time point. For GCaMP6m imaging experiments, a 488-nm laser was used for excitation and GCaMP emission was directed to a photomultiplier tube by means of a DM405/488 dichroic mirror. Images were analysed in ImageJ using the 'measure hyperstack' function. Raw traces were corrected for photobleaching before further analysis. To correct for the effect of a decrease in fluorescence over time due to photobleaching, frames acquired before the application of stimulus (around 60 s before stimulus onset) and frames acquired after the stimulus is removed (around 60 s after stimulus removal) were used to fit to an exponential decay. The fitted curve (baseline) was then subtracted from the raw trace. For each individual trace, we used the formula $\Delta F/F_0$ (%) to calculate changes in fluorescence, $\Delta F = (F_n - F_0)$ where F_n is the intensity of neuron at a given time point calculated as described earlier, and F_0 is the baseline fluorescence value, calculated as the average of intensity over the period when the initial temperature was maintained at 23 °C. Maximum $\Delta F/F_0$ (%) increase in GCaMP6m fluorescence was determined by averaging the maximum percentage increases observed for each trace during the entire duration of the cooling experiment. Maximum $\Delta F/F_0$ (%) decrease in GCaMP6m fluorescence was determined by averaging the maximum percentage decreases observed for each trace over the entire duration of the heating experiment.

Control experiments were performed at constant temperature (23 °C). We also conducted control experiments using eGFP to account for the effects of cooling and heating on the fluorescence intensity of eGFP, because GCaMP6m is a GFP-derived calcium sensor. For aristae-less or body-less experiments, we used fine scissors to remove the aristae or the body 5–10 min before the experiment. Statistical comparisons between treatments and genotypes were performed using ANOVA with Tukey's honest significant difference tests. Unpaired two-tailed Student's *t*-tests were used for two-group comparisons. All plots and statistical tests were generated and performed using OriginPro 2016 64Bit (OriginLab Corporation).

Gradual temperature ramp experimental protocol. Adult male flies (3–5 days old) were placed individually in glass capillary tubes containing 2% agar and 4% sucrose food, which were then loaded into TriKinetics DAM2 *Drosophila* Activity Monitors (Waltham) for locomotor activity recordings. Flies were entrained to temperature cycles with gradual ramping of temperature between 18 °C and 25 °C in constant darkness for 6–7 days, and then released into constant temperature (25 °C) for 7 more days. Zeitgeber time (ZT) 0 marks the beginning of the heating phase (18 °C to 25 °C), and ZT12 marks the onset of the cooling phase (25 °C to 18 °C). The rate of temperature change was approximately 0.6 °C h^{-1}.

Locomotor activity analysis. Beam-crossing counts were organized into 30-min bins for time-series analysis of locomotor activity. Averaged population activity profiles under temperature cycles were generated using a previously described counting macro[33]. In brief, activity levels during temperature cycles were normalized for each fly, such that the average number of the beam crossings in each day (48 bins) is equal to 1. Next, the daily normalized activity profile of individual flies was averaged over the last four days of the temperature cycle to generate an average activity profile for each fly. Finally, the population average of normalized activity was determined, and the results are displayed as normalized activity plots (Fig. 4, Extended Data Figs 5, 8–10).

Sleep analysis. For sleep measurements, beam-crossing counts were organized into 1-min bins. Sleep was defined as bouts of uninterrupted inactivity lasting for five or more minutes[34]. Sleep plots were generated as described previously[33]. Sleep plots displayed in Fig. 4, Extended Data Figs 5, 8–10 indicate the averaged population sleep profiles and were obtained by averaging the sleep data over the last four days of temperature cycles across the entire population.

Transient temperature change protocol. For unexpected cooling experiments, flies were entrained to a 22 °C to 28 °C ramping temperature cycle for a week, and the temperature was decreased from 25 °C to 22 °C for one hour during middle of the heating phase on the 8th day of entrainment (Extended Data Fig. 5e). Flies were entrained to a 18 °C to 28 °C ramping temperature cycle for a week and the temperature was decreased from 25 °C to 18 °C for one hour in the middle of the heating phase on the 8th day of entrainment (Extended Data Fig. 5f). For unexpected heating experiments, flies were entrained to a 22 °C to 28 °C ramping temperature cycle for a week and the temperature was transiently increased from 25 °C to 28 °C for one hour in the middle of the cooling phase on the 8th day of entrainment (Extended Data Fig. 5g).

Rhythmicity analysis. We used activity counts of individual flies under constant temperature conditions after temperature entrainment to analyse rhythmicity and to determine the free-running period of the circadian clock (Extended Data Fig. 6a). Rhythmicity and free-running periods of individual flies were determined by a χ^2 periodogram analysis with a confidence level of 0.01 using the software ClockLab Analysis 6 (Actimetric). The power and significance values generated from the χ^2 analysis were used to calculate 'rhythmic power' as a measure of the strength of each rhythm, as described previously[35].

Anticipation index. The extent to which the daily major increase in sleep anticipates the onset of the cooling phase, termed 'anticipation', was quantified as follows. First, for each individual fly, the sleep data were averaged over the last four days of temperature cycles after entrainment (Extended Data Fig. 10a, b, top panels). Sleep data were plotted as the amount of sleep within 10-min interval bins. For each time point of sleep data, the change in sleep over the subsequent hour was quantified by fitting a linear regression to the sleep data for the following hour, which we call the 'sleep slope' (Extended Data Fig. 10a, b, bottom panels). The sleep slope therefore represents the rate at which sleep is changing over the subsequent hour, with positive slope indicating an increase in sleep and negative slope indicating a decrease in sleep. Next, we identified the local maximum slope value residing closest to ZT12 (the onset of cooling), which we refer to as ZT'A', such that the sleep slopes at ZT'A' and for the next 2 h are non-negative. Therefore, ZT'A' represents the point closest to the onset of cooling at which sleep started to coherently and precipitously increase. Anticipation is calculated as ZT12 − ZT'A' for each individual fly. Anticipation measurements were averaged for all flies of a given genotype and condition and compared to their corresponding controls using a one-way ANOVA and a Tukey's honest significant difference test. For aristae-less *nocte[1]* flies, which are largely arrhythmic under ramping temperature cycles, we were not able to identify a ZT'A' that satisfies all the conditions described above, therefore an anticipation index was not calculated. Positive anticipation indices indicate that the flies increased their sleep before the onset of cooling. Low or negative anticipation index reflects a loss of cooling anticipation (Fig. 4f).

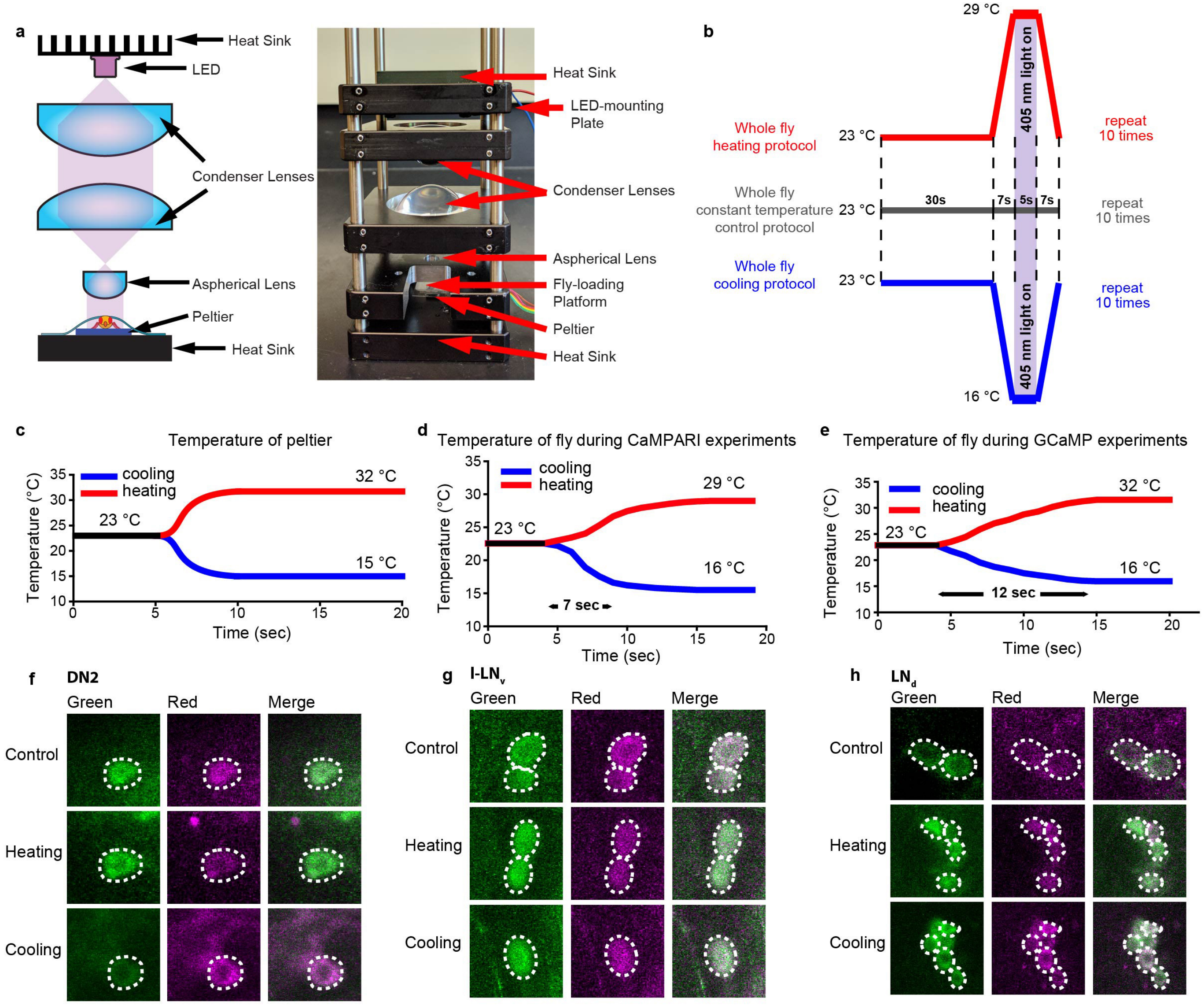
To quantify the extent to which the activity peak anticipates the onset of cooling, we followed a quantitative approach that has been described previously[36]. In brief, we smoothed the raw activity profiles of individual flies for the last four days of the temperature cycles with a low-pass Butterworth filter (Extended Data Fig. 10c, d). The activity peak was identified as the local maximum of the filtered activity profile closest to ZT12 (the onset of cooling) for each individual day. For each individual fly, the activity peak phase was determined by averaging the peak phases across the four days. The average activity peak phase is calculated as the population average for each genotype and compared to their corresponding controls using one-way ANOVA with Tukey's honest significant difference tests (Extended Data Fig. 10n).

Cooling index. A cooling index was used to quantify the precipitous increase in sleep associated with the onset of cooling (Extended Data Fig. 6b). It was calculated as follows: using the average sleep profiles of each individual fly (as described above for our anticipation metric), sleep values during the first hour of the cooling phase (ZT12–ZT13) were fitted to a linear equation and the slope of this line was calculated. The cooling index was obtained by averaging the slopes of the entire population and the experimental genotypes were compared to the corresponding controls using one-way ANOVA with Tukey's honest significant difference tests. A higher sleep index signifies a sharper transition to sleep at the onset of cooling.

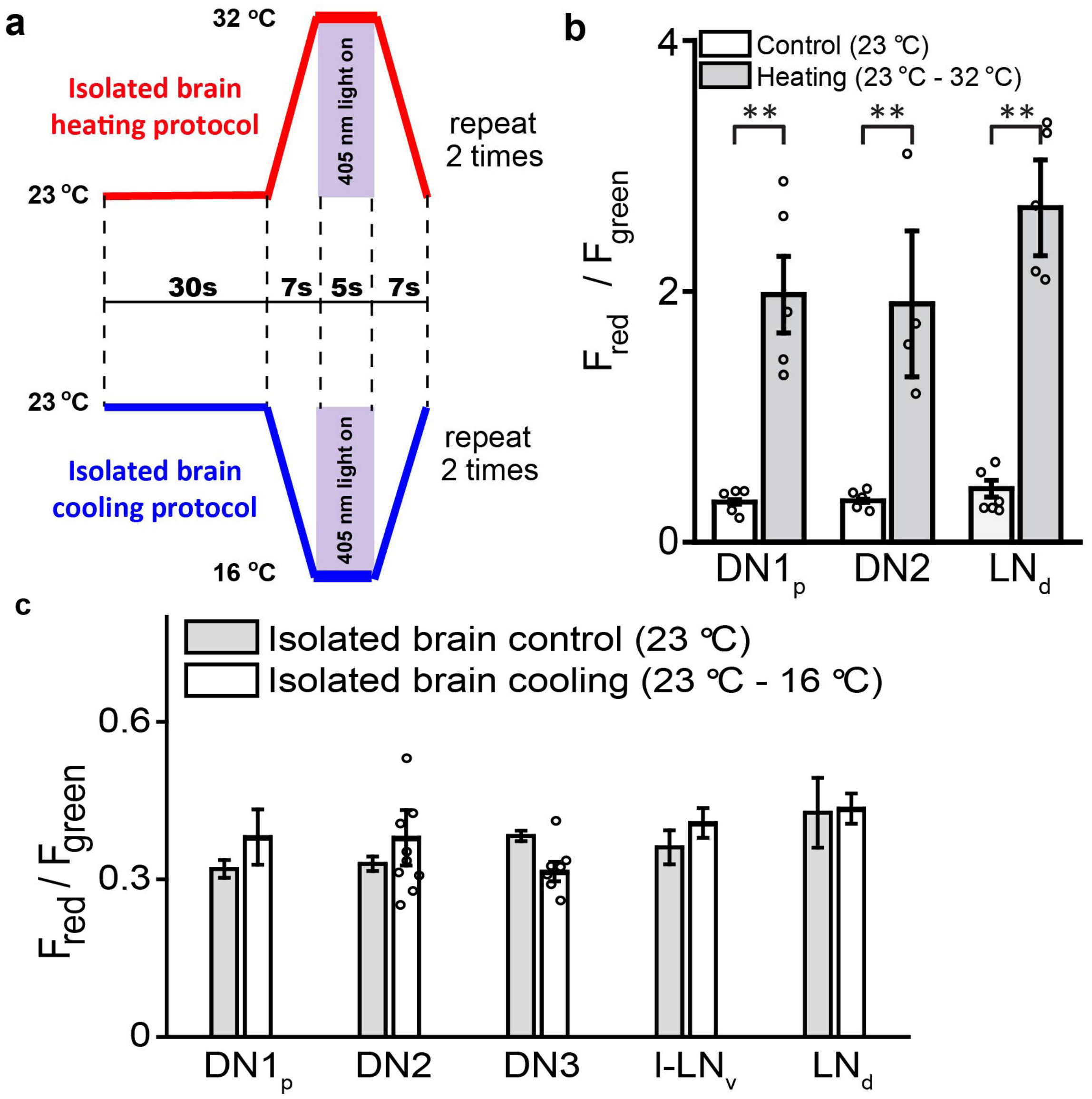
Heating index. The heating index was calculated as the Pearson product-moment correlation coefficient between the activity and the temperature during the heating phase for each individual fly (Extended Data Fig. 6a). The average heating index was obtained by averaging the heating indices of the entire population of flies. We used activity levels between ZT3 and ZT9 to exclude potential startle responses to the onset of heating and anticipation of cooling at the end of the heating phase. A correlation coefficient close to 1 suggests that the locomotor activity and the temperature increase during the heating phase are closely correlated.

Data availability. Source Data for Figs 1–4 and Extended Data Figs 2–4, 6, 7 and 10 are available with the online version of this paper. All other data are available from the corresponding author upon request.

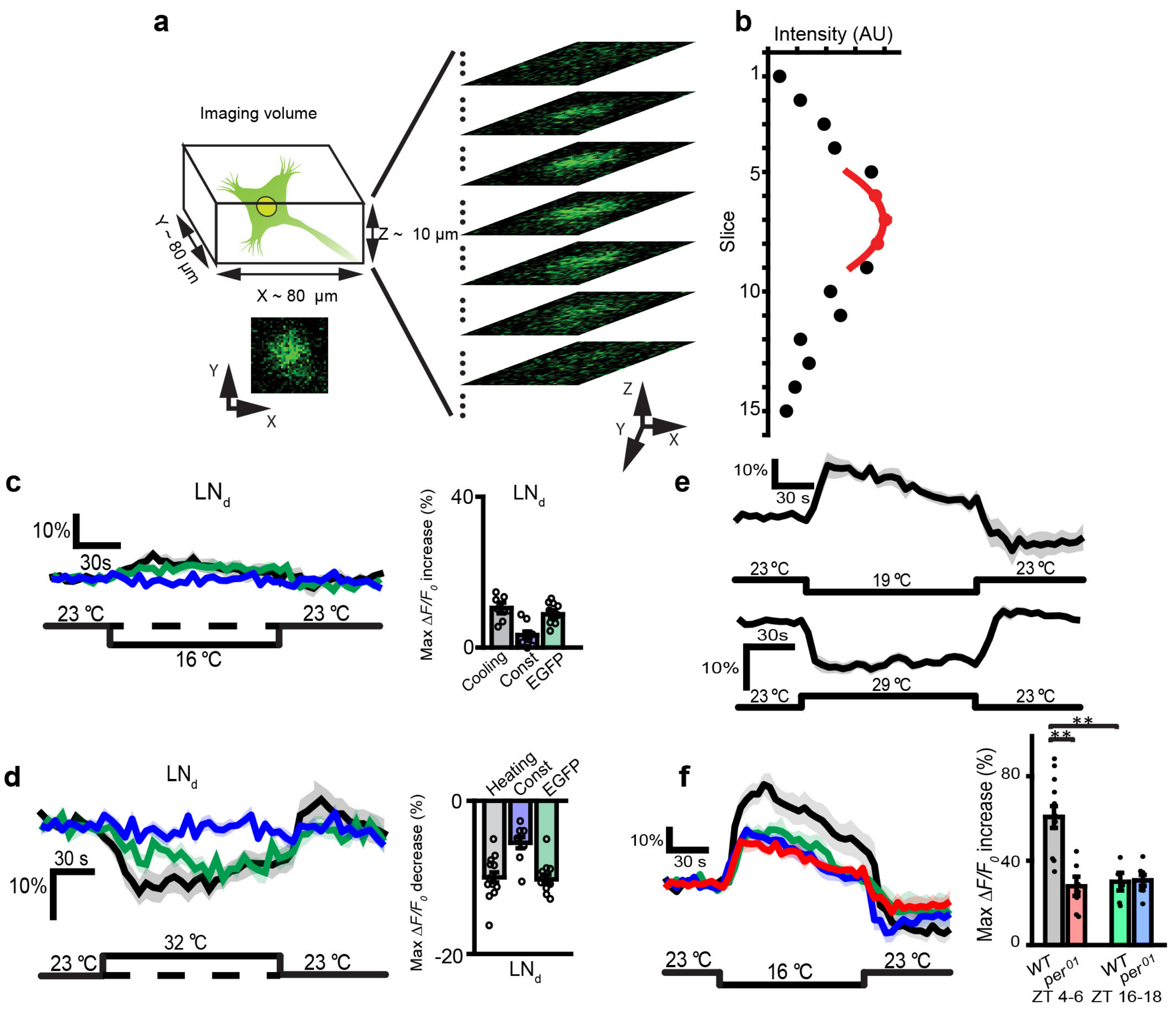
29. Gummadova, J. O., Coutts, G. A. & Glossop, N. R. J. Analysis of the *Drosophila* clock promoter reveals heterogeneity in expression between subgroups of central oscillator cells and identifies a novel enhancer region. *J. Biol. Rhythms* **24,** 353–367 (2009).
30. Konopka, R. J. & Benzer, S. Clock mutants of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **68,** 2112–2116 (1971).
31. Yao, Z., Macara, A. M., Lelito, K. R., Minosyan, T. Y. & Shafer, O. T. Analysis of functional neuronal connectivity in the *Drosophila* brain. *J. Neurophysiol.* **108,** 684–696 (2012).
32. Liang, X., Holy, T. E. & Taghert, P. H. A series of suppressive signals within the *Drosophila* circadian neural circuit generates sequential daily outputs. *Neuron* **94,** 1173–1189.e4 (2017).
33. Pfeiffenberger, C., Lear, B. C., Keegan, K. P. & Allada, R. Processing circadian data collected from the Drosophila Activity Monitoring (DAM) System. *Cold Spring Harb. Protoc.* https://doi.org/10.1101/pdb.prot5519 (2010).
34. Hendricks, J. C. *et al.* Rest in *Drosophila* is a sleep-like state. *Neuron* **25,** 129–138 (2000).
35. Yao, Z. & Shafer, O. T. The *Drosophila* circadian clock is a variably coupled network of multiple peptidergic units. *Science* **343,** 1516–1520 (2014).
36. Yao, Z., Bennett, A. J., Clem, J. L. & Shafer, O. T. The *Drosophila* clock neuron network features diverse coupling modes and requires network-wide coherence for robust circadian rhythms. *Cell Rep.* **17,** 2873–2881 (2016).



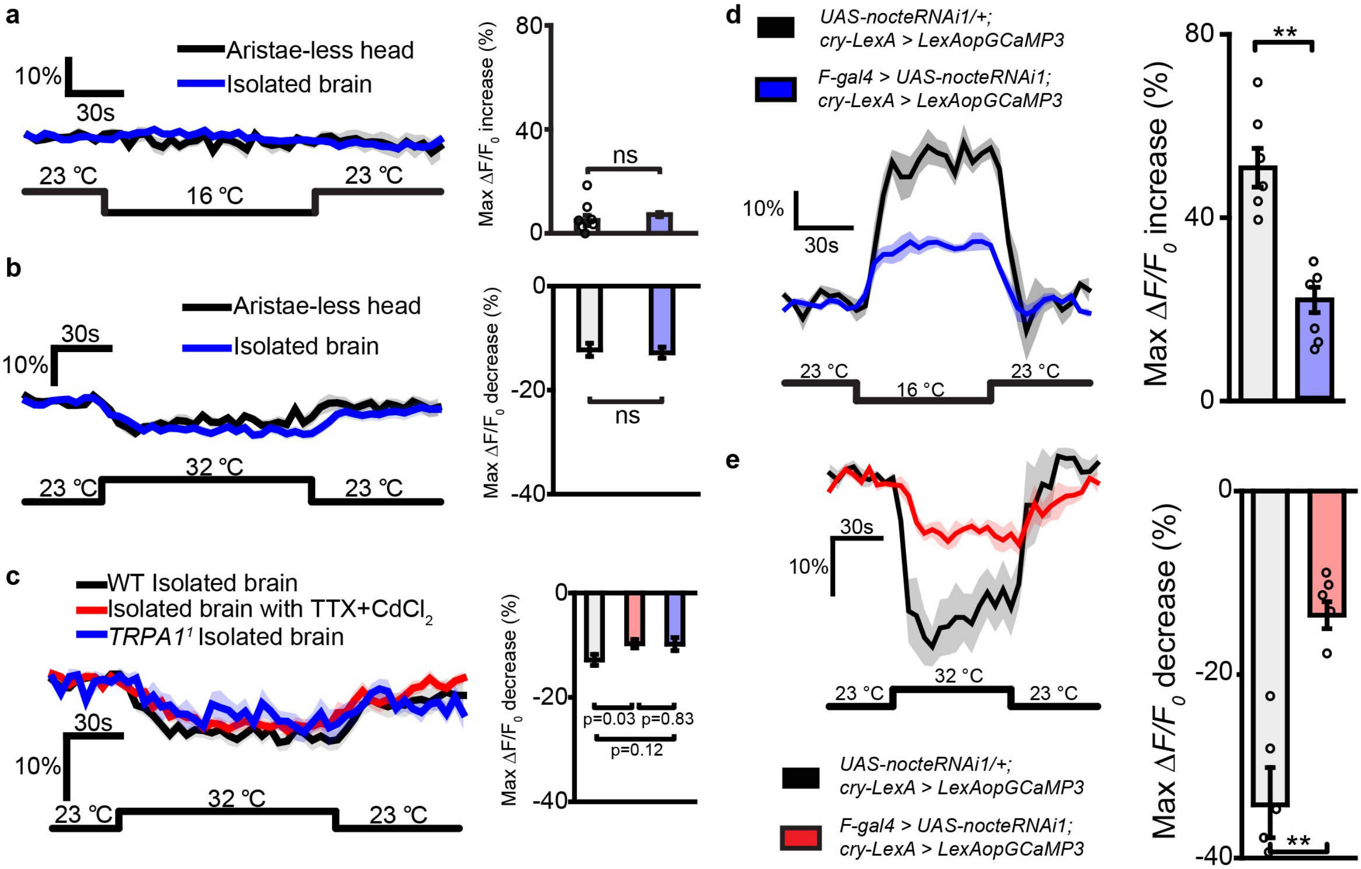
Extended Data Figure 1 | CaMPARI protocol for the identification of temperature-sensitive clock neurons. a, Schematic (left) and photograph (right) of the illumination and thermal control system for the CaMPARI experiments (see Methods). **b**, Temperature and LED illumination protocol for whole-fly CaMPARI experiments (see Methods). **c**–**e**, Measured temperature of the Peltier element (**c**), and the fly during CaMPARI (**d**) and GCAMP (**e**) experiments. Heating and cooling stimuli were provided by changing the Peltier element set-point from 23 °C to 32 °C (red line) and from 23 °C to 15 °C (blue line), respectively (see Methods). **f**–**h**, Representative confocal microscopy images of green and red CaMPARI fluorescence in the DN2s (**f**), l-LN$_v$s (**g**), and LN$_d$s (**h**) in response to constant temperature (23 °C, top panel), heating (23 °C to 29 °C, middle panel), and cooling (23 °C to 16 °C, bottom panel). These experiments were repeated independently using seven brains with similar results.



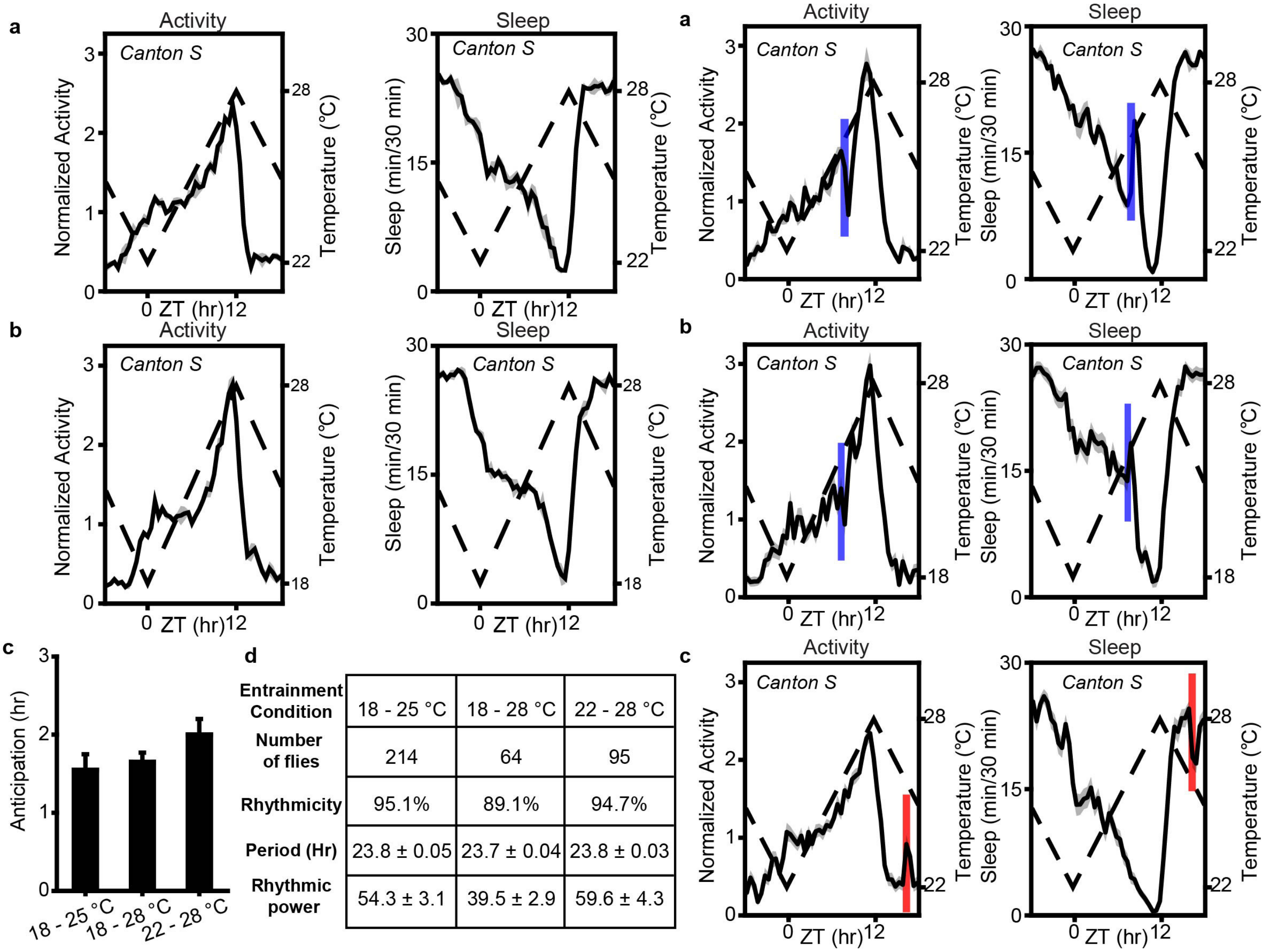
Extended Data Figure 2 | CaMPARI photoconversion in isolated brains. **a**, Temperature stimuli and LED illumination protocol for CaMPARI experiments on isolated brains. In these experiments, the temperature was increased from 23 °C to 32 °C and only two cycles of illumination were used to avoid excessive photoconversion (see Methods). **b**, Quantification of the ratio of red and green CaMPARI fluorescence in clock neurons from isolated brains ectopically expressing the thermosensitive channel TRPA1[20], which is activated by temperatures greater than 25 °C. Clock neurons ectopically expressing TRPA1 display high levels of CaMPARI photoconversion in response to heating (grey bars, number of neurons: 5 DN1$_p$s, 4 DN2s, 5 LN$_d$s) compared to constant temperature controls (white bars, number of neurons: 5 DN1$_p$s, 5 DN2s, 6 LN$_d$s). **c**, Ratio of red to green CaMPARI fluorescence in clock neurons in isolated brains in response to cooling (white bars, 63 DN1$_p$s, 9 DN2s, 7 DN3s, 19 LN$_d$s, 20 l-LN$_v$s) and constant temperature (grey bars, 33 DN1$_p$s, 10 DN2s, 10 DN3s, 26 LN$_d$s, 23 l-LN$_v$s). Histograms represent the mean ratio of red to green fluorescence ± s.e.m., three to four brains analysed for each condition. Individual data points are shown, as the sample size is less than 10. $^{**}P < 0.005$, unpaired two-tailed Student's *t*-test. Specific *P* values are reported in the Source Data for this figure.



Extended Data Figure 3 | GCaMP protocol for the quantification of clock neuron responses to temperature changes. a, **b**, Schematic of the GCaMP volume imaging (**a**) and quantification (**b**) during temperature changes (see Methods). **c**, Averaged LN$_d$ GCaMP6m fluorescence changes during cooling (black, *n* = 8 neurons) and at constant temperature (blue, *n* = 7). Averaged eGFP fluorescence traces from LN$_d$s during cooling (green, *n* = 9). These data were summarized as average maximum $\Delta F/F_0$ (%) increase ± s.e.m. in the histograms (right). LN$_d$s displayed no apparent GCaMP responses to cooling. **d**, Averaged LN$_d$ GCaMP6m fluorescence traces during heating (black, *n* = 12). Constant temperature GCaMP6m traces (blue, same data as in **c**) and eGFP traces during heating (green, *n* = 10 neurons) are also shown. These data were summarized as average minimum $\Delta F/F_0$ (%) ± s.e.m. in the histograms (right). The LN$_d$s appeared to display a loss of GCaMP fluorescence in response to heating; however, eGFP fluorescence displayed a similar response to heating (unpaired two-tailed Student's *t*-test), revealing this response to be an artefact. **e**, Average DN1$_p$ GCaMP6m responses are proportional to the magnitude of the cold stimulus (top, *n* = 8) or hot stimulus (bottom, *n* = 26). **f**, Averaged DN1$_p$ GCaMP6m responses from wild-type (WT) flies at two different times during the diurnal cycle (black, ZT4–6, *n* = 10; green, ZT16–18, *n* = 5), and *per^{01}* flies at the same time points (red, ZT4–6, *n* = 7; blue, ZT16–18, *n* = 8). For the averaged GCaMP6m fluorescence traces, dark lines indicate the mean and shaded areas indicate the s.e.m. *n* = the number of LN$_d$s as stated in **c**, **d** and *n* = the number of DN1$_p$s as stated in **e**, **f**. Individual data points are shown, as *n* is less than 10. Five to seven brains were analysed in each experiment. Temperature plots are displayed below the GCaMP traces. Specific *P* values are reported in the Source Data for this figure.



Extended Data Figure 4 | DN1$_p$ GCaMP responses in isolated brains, and in flies in which *nocte* is downregulated in the chordotonal organs. **a, b**, Averaged DN1$_p$ GCaMP6m fluorescence traces during cooling (**a**) and heating (**b**), after flies were decapitated and aristae were removed, leaving the rest of the antennae and head cuticle intact. Data in the histograms are presented as average maximum increase (**a**) and decrease (**b**) of $\Delta F/F_0$ (%) ± s.e.m. for cooling and heating, respectively. Responses were not significantly different between heads from which aristae were removed (black, $n = 7$ in **a**, $n = 11$ in **b**) and isolated brains (blue, $n = 32$ in **a**, $n = 25$ in **b**). These data suggest that aristae are the only peripheral thermoreceptors on the head that contribute to DN1$_p$ temperature responses. The thermoreceptors of the sacculi, which remained after arista removal, are not sufficient for DN1$_p$ temperature responses, although it is possible that the cutting of the aristae may have had non-specific effects on saccular function. Unpaired two-tailed Student's *t*-test. **c**, GCaMP6m responses of DN1$_p$s in response to heating in isolated brains in HL3 solution (black, $n = 25$) and in HL3 supplemented with 2 μM tetrodotoxin and 400 μM CdCl$_2$ to prevent neuronal firing and synaptic transmission (red, $n = 20$), and in the explanted brains of *TRPA1*^{1} flies in HL3 (blue, $n = 11$). Histogram plotted as in **b**, ANOVA was performed across the three conditions ($F_{2,55} = 1.395$, $P > 0.05$) and Tukey's honest significant difference tests were conducted in pairwise fashion with *P* values indicated on the histogram (right). These results suggest that the loss of GCaMP fluorescence seen in DN1$_p$s in isolated brains in the absence of aristae and chordotonal organs is a non-physiological artefact. **d**, Averaged GCaMP3 fluorescence traces from DN1$_p$s during cooling in control *UAS-nocteRNAi1/+* flies (black, $n = 6$) and *F-gal4>UAS-nocteRNAi1* flies (blue, $n = 6$) (*F-Gal4* drives expression predominantly in chordotonal organs[7]). Histogram plotted as in **a**. **e**, Averaged GCaMP3 fluorescence traces from the DN1$_p$s during heating in control *UAS-nocteRNAi1/+* (black, $n = 5$) and *F-gal4>UAS-nocte RNAi1* flies (red, $n = 5$). Histogram plotted as in **b**. For averaged GCaMP fluorescence traces, dark lines indicate mean and shaded areas indicate s.e.m. n = number of DN1$_p$s, individual data points shown when n is less than 10. Four to six brains were analysed for each condition. $^{**}P < 0.005$, unpaired two-tailed Student's *t*-test. Individual *P* values are reported in the Source Data for this figure.



Entrainment Condition	18 - 25 °C	18 - 28 °C	22 - 28 °C
Number of flies	214	64	95
Rhythmicity	95.1%	89.1%	94.7%
Period (Hr)	23.8 ± 0.05	23.7 ± 0.04	23.8 ± 0.03
Rhythmic power	54.3 ± 3.1	39.5 ± 2.9	59.6 ± 4.3

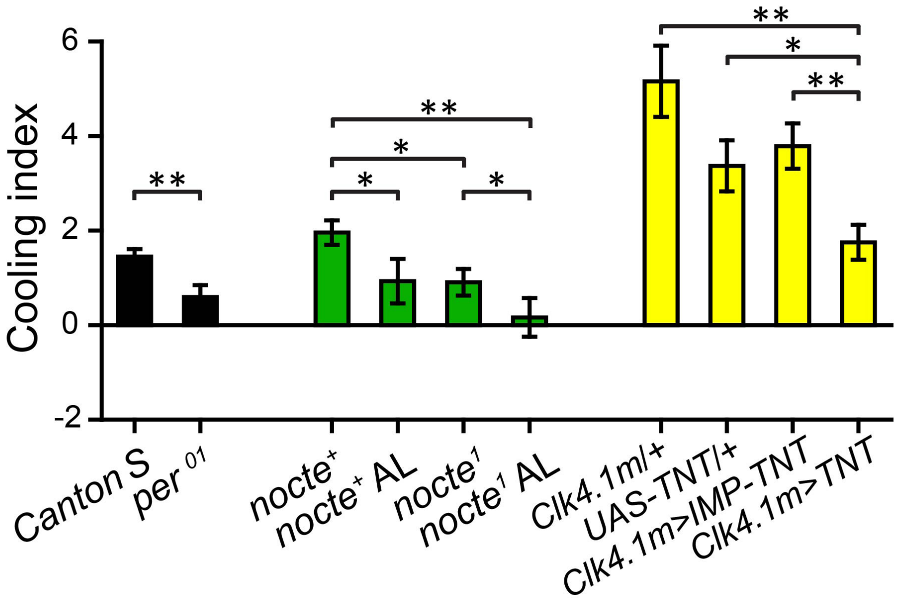
Extended Data Figure 5 | Activity and sleep profiles of wild-type *Canton S* flies under different amplitudes of ramping temperature cycles. **a**–**g**, *Canton S* flies were entrained to ramping temperature cycles of differing amplitudes and ranges: 18 °C to 28 °C or 22 °C to 28 °C for a week, followed by release into constant temperature conditions (25 °C) for another week. **a**, **b**, Averaged population activity of flies over the last four days of the temperature cycles (left) and averaged sleep plots under the same conditions (right). Dark lines indicate mean and shaded regions indicate s.e.m. *n* values are shown in **d**. **c**, Quantification of the anticipation of cooling for flies entrained to ramping temperature cycles from 18 °C to 25 °C, 18 °C to 28 °C and 22 °C to 28 °C. Data represent mean ± s.e.m. *n* values are shown in **d**. **d**, Summary of free-running locomotor activity rhythms of *Canton S* flies under constant conditions after entrainment to ramping temperature cycles. Data represent population average ± s.e.m. **e**–**g**, Averaged population locomotor activity on the day in which the temperature was unpredictably decreased (**e**, **f**) or increased (**g**) (left) and averaged sleep profiles (right) (see Methods) (*n* = 32 (**e**), *n* = 61 (**f**), *n* = 63 (**g**)). Blue bars indicate unanticipated cooling, red bars indicate unanticipated heating. Dark lines indicate mean and shaded regions indicate s.e.m. Black dashed lines in the plots represent temperature. *n* = number of flies.

a

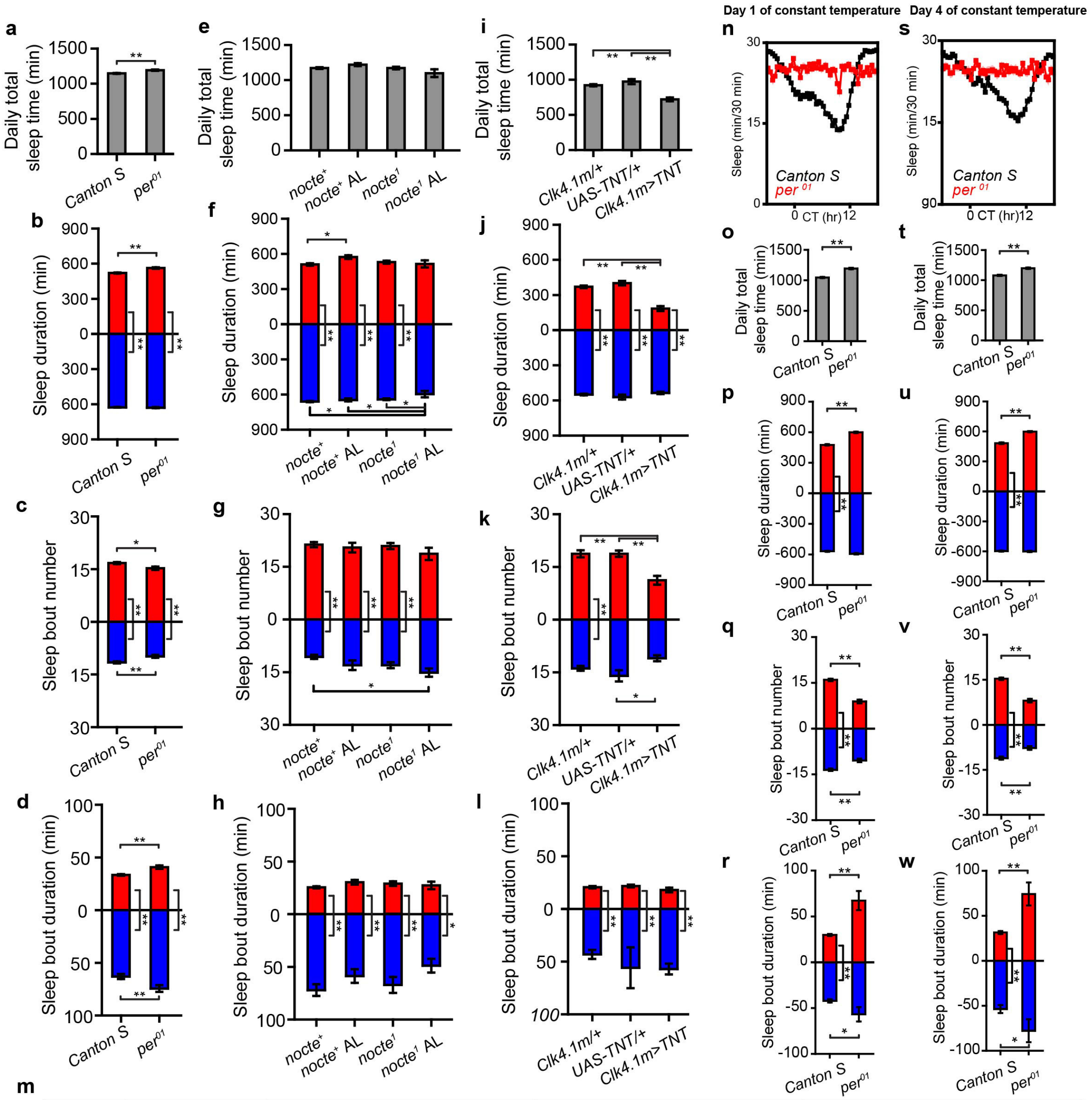
Genotype	Number of flies	Heating index	Rhythmicity	Period (Hr)	Rhythmic power	WT-like behavior
Canton S	214	0.92	95.1%	23.82 ± 0.05	54.3 ± 3.1	Yes
per^{01}	96	-0.60	7.4%	29.93 ± 1.41	1.1 ± 0.3	No
$Nocte^{+}$	46	0.95	87.0%	23.69 ± 0.09	29.5 ± 3.8	Yes
***$Nocte^{+}$* AL**	28	0.25	51.5%	23.59 ± 0.09	13.9 ± 2.4	No
$Nocte^{1}$	56	0.84	39.3%	24.93 ± 0.35	8.4 ± 1.8	No
***$Nocte^{1}$* AL**	44	0.31	13.5%	22.82 ± 3.62	3.9 ± 2.0	No
Clk4.1m/+	17	0.85	80.3%	24.32 ± 0.31	17.8 ± 4.1	Yes
UAS-TNT/+	23	0.86	85.2%	23.70 ± 0.12	16.1 ± 2.4	Yes
Clk4.1m>IMP-TNT	32	0.96	89.7%	23.58 ± 0.07	34.2 ± 5.3	Yes
Clk4.1m>TNT	100	0.49	86.8%	23.95 ± 0.08	36.4 ± 4.2	No
DN2>IMP-TNT	24	0.94	95.0%	23.87 ± 0.09	33.6 ± 5.3	Yes
DN2>TNT	40	0.94	94.4%	23.91 ± 0.32	36.5 ± 3.9	Yes

Constant Temperature

b



Extended Data Figure 6 | Locomotor behaviour of wild-type and genetically manipulated flies under ramping temperature cycles and under constant conditions after entrainment. a, Summary of heating index under ramping temperature cycles and free-running locomotor activity rhythms under constant conditions after entrainment to 18 °C to 25 °C ramping temperature cycles. *per^{01}*, aristae-less *$nocte^{+}$* flies (*$nocte^{+}$* AL), aristae-less *$nocte^{1}$* mutants (*$nocte^{1}$* AL) and *Clk4.1m>UAS-TNT* flies, in which synaptic transmission is blocked in DN1$_{p}$s, displayed strong behavioural phenotypes during the heating phase of the temperature cycle as revealed by low correlation coefficients between activity and temperature during the heating phase (see Methods). *per^{01}*, *$nocte^{1}$* mutants and flies lacking aristae displayed arrhythmic locomotor activity under constant conditions after entrainment to ramping temperature cycles. Data were presented as population average ± s.e.m. **b**, A summary of cooling indices (see Methods) for experimental and control flies. *n* values are shown in **a**. Data were presented as average of cooling index ± s.e.m. A one-way ANOVA was conducted and Tukey's honest significant difference tests were used; *$P < 0.05$, **$P < 0.005$. n = number of flies. Individual *P* values are reported in the Source Data for this figure.

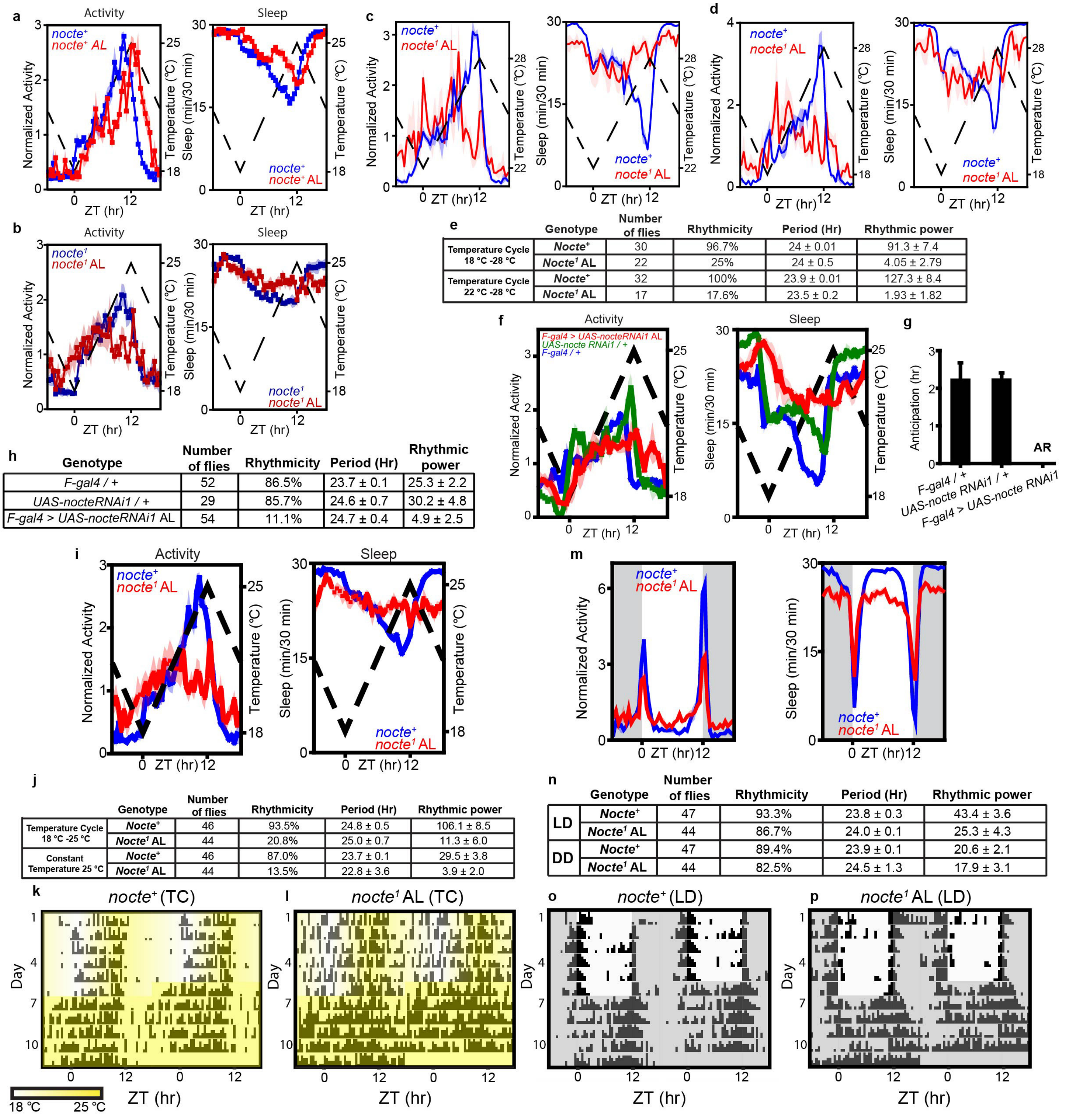


m

Panel	Total Sleep time	Sleep time		Sleep bout number		Sleep bout duration	
		Heating	Cooling	Heating	Cooling	Heating	Cooling
a-d	$F_{1,309}$=15.9, p=8.5889E-5	$F_{1,309}$=28.1, p=2.25658E-7	$F_{1,309}$=0.4	$F_{1,309}$=7.8, p=0.00547	$F_{1,309}$=12.3, p=5.21632 E-4	$F_{1,309}$=21.0, p=6.7765E-6	$F_{1,309}$=7.33, p=0.00716
e-h	$F_{3,155}$=1.8	$F_{3,155}$=2.5	$F_{3,155}$=5.6, p=0.00124	$F_{3,155}$=0.8	$F_{3,155}$=3.4, p=0.02009	$F_{3,155}$=1.0	$F_{3,155}$=1.4
i-l	$F_{2,139}$=26.0, p=2.80163E-11	$F_{2,139}$=42.2, p=7.21645E-15	$F_{2,139}$=2.1	$F_{2,139}$=15.0, p=6.30466E-6	$F_{2,139}$=6.4, p=0.00311	$F_{2,139}$=1.11	$F_{2,139}$=0.78

Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Sleep characteristics of wild-type and genetically manipulated flies under ramping temperature cycles and constant conditions after entrainment to ramping temperature cycles. **a**–**l**, Sleep data analysis for *Canton S* and genetically modified flies under 18 °C to 25 °C temperature ramping cycles (see Methods). Daily total sleep time (**a**, **e**, **i**); total sleep time during heating (red) and cooling (blue) phases (**b**, **f**, **j**); sleep bout number during heating (red) and cooling (blue) phases (**c**, **g**, **k**); and sleep bout duration during heating (red) and cooling (blue) phases (**d**, **h**, **l**) of the genotypes indicated. Number of flies used in the analysis: **a**–**d**, *Canton S* ($n = 214$), per^{01} ($n = 96$); **e**–**h**, $nocte^{+}$ ($n = 40$), $nocte^{+}$ AL ($n = 28$), $nocte^{1}$ ($n = 49$) and $nocte^{1}$ AL ($n = 39$); **i**–**l**, *Clk4.1m/+* ($n = 17$), *UAS-TNT/+* ($n = 23$) and *Clk4.1m>TNT* ($n = 100$). Data are shown as population average ± s.e.m. n = number of flies. **m**, Statistical analysis. For determining statistical significance, ANOVA was performed across each individual group. Tukey's honest significant difference tests were conducted in pairwise fashion with *P* values indicated on the graphs: $^{**}P < 0.005$, $^{*}P < 0.05$. Unpaired two-tailed Student's *t*-tests were conducted for two-group comparison of the heating phase and the cooling phase within each individual genotype. The number of flies is shown in panels **a**–**l**. **n**, Averaged population sleep plots on day 1 of constant conditions after entrainment to ramping temperature cycles for wild-type *Canton S* (black, $n = 198$) and mutant per^{01} (red, $n = 87$) flies. Circadian time (CT) 0 is the start of the subjective heating phase and CT12 is the start of the subjective cooling phase, the times when the temperature transitions would have occurred had the temperature cycle continued. Dark lines indicate mean and shaded regions in the plots indicate s.e.m. **s**, Averaged population sleep plots on day 4 of constant temperature after entrainment for *Canton S* (black, $n = 198$) and per^{01} (red, $n = 83$) flies. Dark lines indicate mean and shaded regions in the plots indicate s.e.m. **o**–**r**, **t**–**w**, Daily total sleep time, sleep times during subjective heating (red) and cooling (blue) phases, sleep bout number during subjective heating (red) and cooling (blue) phases, and sleep bout duration during subjective heating (red) and cooling (blue) phases of *Canton S* and per^{01} on day 1 (**o**–**r**) and day 4 (**t**–**w**) of constant conditions. The number of flies in panels **o**–**r** is shown in **n**, and the number of flies in panels **t**–**w** is shown in **s**. Data are presented as mean ± s.e.m. *Canton S* flies cycles under constant conditions display similar sleep patterns as observed during temperature cycles. By contrast, per^{01} flies do not display any differences in their sleep time or bout number or bout duration during the subjective heating phase and cooling phase, consistent with a lack of circadian timekeeping. n = number of flies. Statistical analysis was performed using unpaired two-tailed Student's *t*-test. $^{**}P < 0.005$, $^{*}P < 0.05$. Individual *P* values are reported in the Source Data for this figure. Analysis of the sleep data revealed the following: 1. Wild-type flies sleep significantly more during the cooling phase than during the heating phase. 2. Wild-type flies have significantly more sleep bouts during the heating phase than during the cooling phase, indicating that sleep is more fragmented during the heating phase. 3. Wild-type flies have significantly longer sleep bout durations during the cooling phase compared to during the heating phase, indicating that sleep is more consolidated during the cooling phase. 4. Manipulation of chordotonal organ or aristae function does not produce changes in total sleep or sleep quality, only in the timing of sleep. 5. $nocte^{1}$ mutant flies that lack aristae ($nocte^{1}$ AL) fail to show significant differences in the amount of sleep during the heating and cooling phases. 6. Inhibition of a subset of $DN1_{p}$s results in decreased total daily sleep time and specifically reduces sleep time during the heating phase.



e

Genotype		Number of flies	Rhythmicity	Period (Hr)	Rhythmic power
Temperature Cycle 18 °C -28 °C	*Nocte*$^{+}$	30	96.7%	24 ± 0.01	91.3 ± 7.4
	Nocte^{1} AL	22	25%	24 ± 0.5	4.05 ± 2.79
Temperature Cycle 22 °C -28 °C	*Nocte*$^{+}$	32	100%	23.9 ± 0.01	127.3 ± 8.4
	Nocte^{1} AL	17	17.6%	23.5 ± 0.2	1.93 ± 1.82

h

Genotype	Number of flies	Rhythmicity	Period (Hr)	Rhythmic power
F-gal4 / +	52	86.5%	23.7 ± 0.1	25.3 ± 2.2
UAS-nocteRNAi1 / +	29	85.7%	24.6 ± 0.7	30.2 ± 4.8
F-gal4 > UAS-nocteRNAi1 AL	54	11.1%	24.7 ± 0.4	4.9 ± 2.5

j

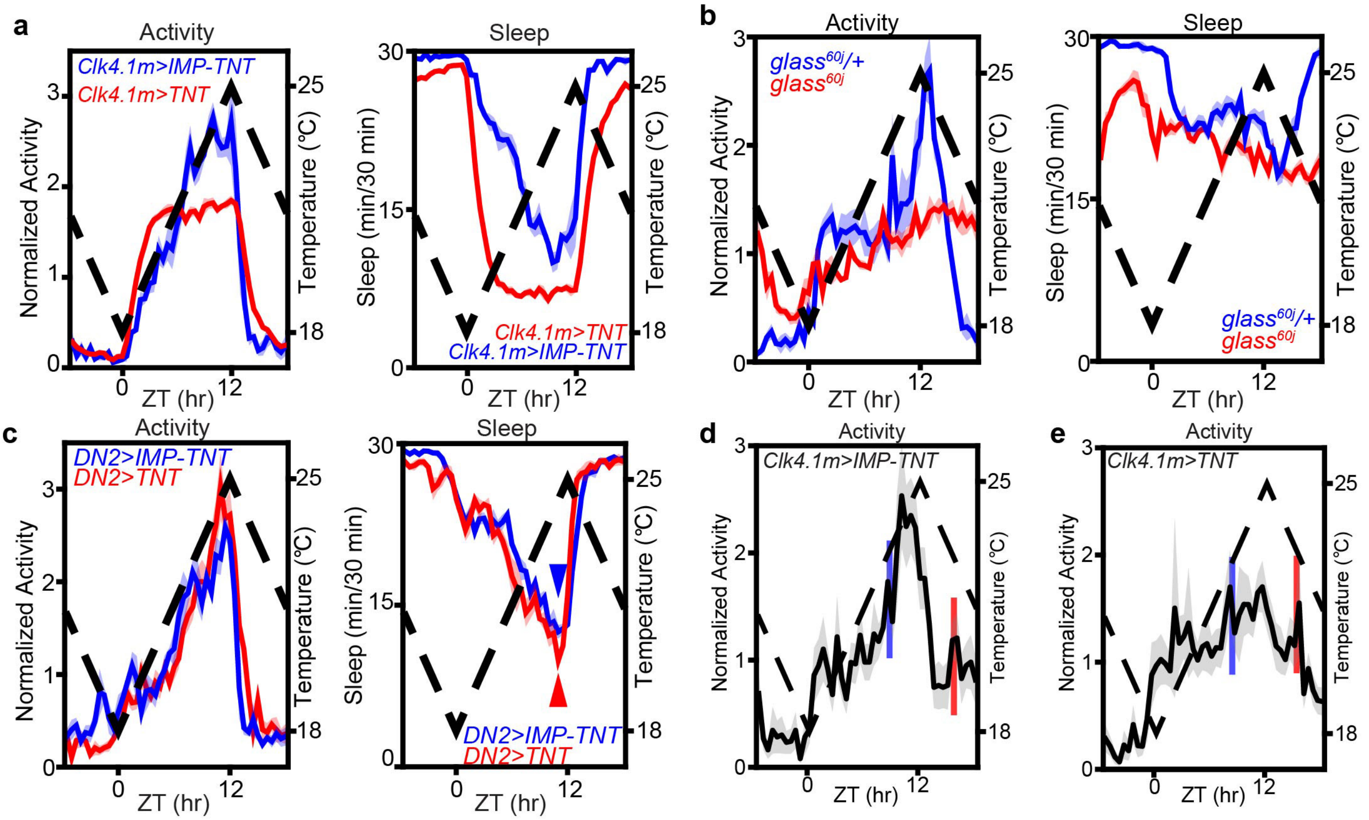
Genotype		Number of flies	Rhythmicity	Period (Hr)	Rhythmic power
Temperature Cycle 18 °C -25 °C	*Nocte*$^{+}$	46	93.5%	24.8 ± 0.5	106.1 ± 8.5
	Nocte^{1} AL	44	20.8%	25.0 ± 0.7	11.3 ± 6.0
Constant Temperature 25 °C	*Nocte*$^{+}$	46	87.0%	23.7 ± 0.1	29.5 ± 3.8
	Nocte^{1} AL	44	13.5%	22.8 ± 3.6	3.9 ± 2.0

n

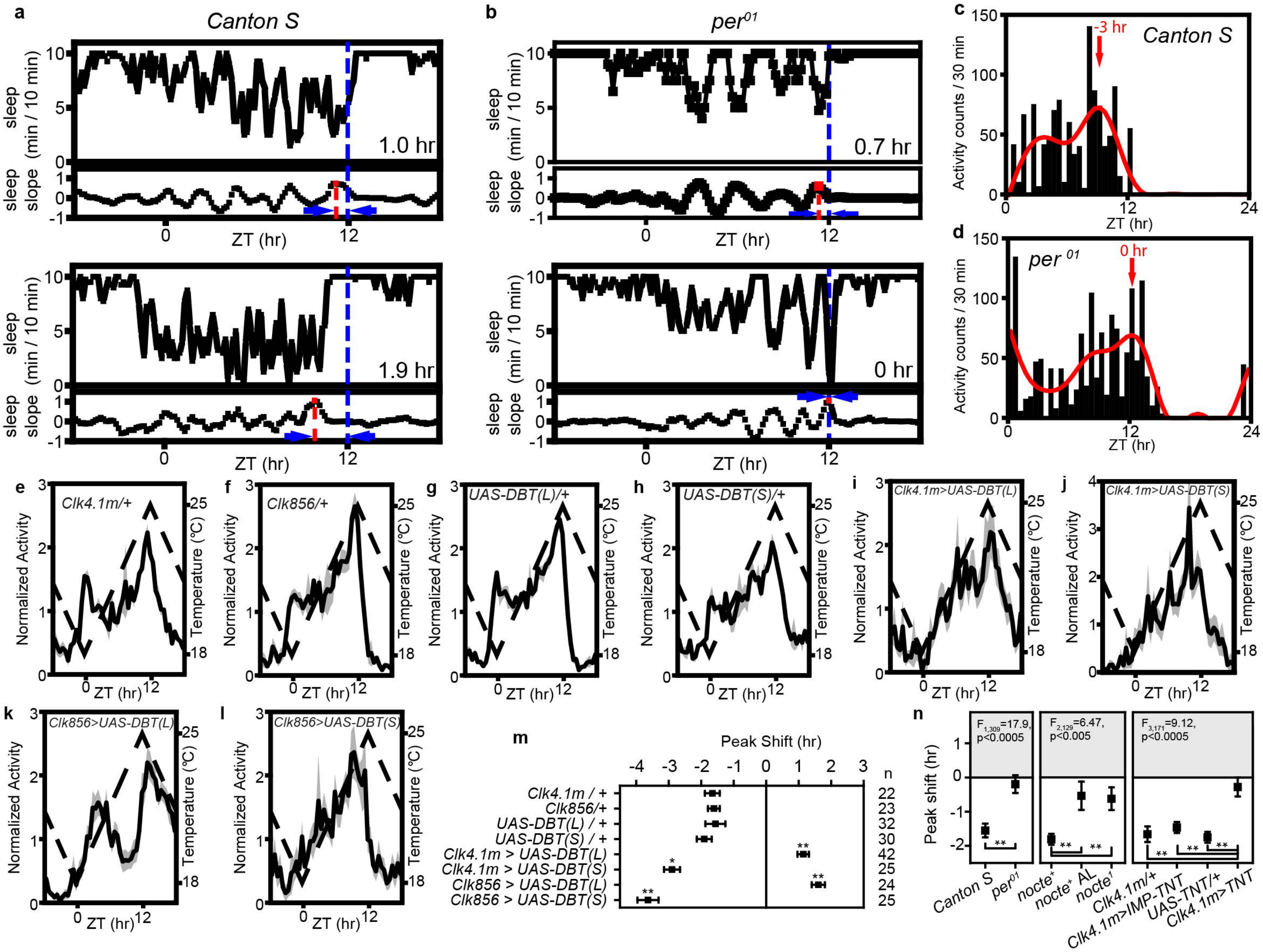
Genotype		Number of flies	Rhythmicity	Period (Hr)	Rhythmic power
LD	*Nocte*$^{+}$	47	93.3%	23.8 ± 0.3	43.4 ± 3.6
	Nocte^{1} AL	44	86.7%	24.0 ± 0.1	25.3 ± 4.3
DD	*Nocte*$^{+}$	47	89.4%	23.9 ± 0.1	20.6 ± 2.1
	Nocte^{1} AL	44	82.5%	24.5 ± 1.3	17.9 ± 3.1

Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Activity and sleep profiles of chordotonal organ mutants and flies lacking aristae. a, b, Locomotor activity averaged over four days of a ramping temperature cycle (left) and averaged sleep profiles (right) for the following flies: **a**, *nocte*$^+$ control flies (blue, $n=46$) and *nocte*$^+$ flies that lack aristae (*nocte*$^+$ AL, red, $n=28$); **b**, *nocte*^{1} mutants (blue, $n=56$), and *nocte*^{1} mutants that lack aristae (*nocte*^{1} AL, red, $n=44$). **c, d**, Locomotor activity profiles averaged over four days of a 22 °C to 28 °C (**c**) and a 18 °C to 28 °C (**d**) ramping temperature cycle for *nocte*$^+$ flies (blue) and *nocte*^{1}*nocte*^{1} mutants that lack aristae (*nocte*^{1} AL, red) (left); averaged sleep profiles (right). The number of flies in **c** and **d** are shown in **e**. **e**, Summary of locomotor activity rhythms under the two different temperature cycling conditions for the genotypes indicated. Data were presented as population average ± s.e.m. **f**, Locomotor activity averaged over four days of a ramping temperature cycle are shown on the left and averaged sleep profiles are shown on the right. The genotypes tested were *F-gal4*/+ (blue), *UAS-nocteRNAi1*/+ (green) and *F-gal4*>*UAS-nocteRNAi1*-AL (red), where AL refers to the aristae-less condition. The number of flies is shown in **h**. **g**, Anticipation of the onset of cooling in the genotypes indicated. *F-gal4*>*UAS-nocteRNAi1*-AL flies are largely arrhythmic (AR), so the anticipation index was not calculated (see Methods). Data are population average ± s.e.m. *n* values are shown in panel **h**. **h**, Summary of locomotor activity rhythms under constant conditions after entrainment. Data are mean ± s.e.m. **i**, Locomotor activity averaged over four days of a 18 °C to 25 °C ramping temperature cycle (left); averaged sleep profiles (right). *nocte*$^+$ control flies are shown in blue and *nocte*^{1} mutants that lack aristae (*nocte*^{1} AL) are shown in red. The number of flies is shown **j**. **j**, Summary of locomotor activity rhythms under a ramping 18 °C to 25 °C temperature cycle and under constant conditions after entrainment for *nocte*$^+$ flies and *nocte*^{1} AL mutants. Data are mean ± s.e.m. **k**, **l**, Double-plotted actograms of a representative *nocte*$^+$ fly (**k**) and *nocte*^{1} AL fly (**l**), which were entrained under temperature ramping cycles for 6 days and released into constant temperature (25 °C). The temperature gradient is shown at the bottom of the actogram. **m**, Locomotor activity averaged over four days of a 12 h:12 h light:dark cycle (left) and averaged sleep profiles (right) for *nocte*$^+$ flies (blue) and *nocte*^{1} AL flies (red). The number of flies is shown in **n**. Lights are turned on at ZT0 and turned off at ZT12. **n**, Summary of locomotor activity rhythms under light:dark (LD) and subsequent constant dark:dark (DD) conditions for *nocte*$^+$ flies and *nocte*^{1} AL mutants. Data are mean ± s.e.m. **o**, **p**, Double-plotted actograms of a representative *nocte*$^+$ fly (**o**) and *nocte*^{1} AL fly (**p**), which were entrained to 12 h:12 h light:dark cycles for 6 days and released into constant darkness. In all activity and sleep plots, dark lines indicate mean and shaded regions indicate s.e.m. Black dashed lines in the plots represent temperature.



Extended Data Figure 9 | Activity and sleep profiles of flies in which synaptic transmission is blocked in DN1$_p$ or DN2 clock neurons. **a**–**c**, Locomotor activity averaged over four days of a ramping temperature cycle (left) and averaged sleep profiles (right) for flies expressing tetanus toxin (TNT) in a subset of the DN1$_p$s, *DN1$_p$>UAS-TNT* experimental flies (red, $n = 100$), and *DN1$_p$>UAS-IMP-TNT* controls expressing inactive TNT control (blue, $n = 32$) (**a**); *glass^{60j}* mutants which lack all DN1$_p$s (red, $n = 49$). and *glass^{60j}/+* heterozygotes (blue, $n = 26$) (**b**), and flies expressing TNT in the DN2s, *DN2>TNT* (red, $n = 40$) and *DN2>IMP-TNT* control (blue, $n = 24$) (**c**). Expression of TNT in DN2s did not affect the locomotor or sleep behaviour of flies under ramping temperature cycles under constant darkness. These results are consistent with previous findings that showed that DN2 synaptic output is required for behaviour synchronization in temperature cycles under constant light conditions, but not under constant dark conditions[8]. See Methods for detailed description of the genetic tools used. **d**, **e**, Averaged population locomotor activity on the day in which the temperature is transiently decreased or increased is shown for *Clk4.1m>IMP-TNT* control flies ($n = 24$) (**d**) and *Clk4.1m>TNT* flies ($n = 20$) (**e**). Flies were entrained to a 18 °C to 25 °C ramping temperature cycle for a week, and the temperature was decreased from 23 °C to 18 °C for an hour during the middle of the heating phase (blue bar) and increased from 22 °C to 25 °C for an hour in the middle of the cooling phase (red bar) on the eighth day of entrainment. Both the control and experimental flies show normal behavioural responses to unexpected changes in temperature, suggesting that blocking synaptic transmission from *Clk4.1m* expressing DN1$_p$s does not have an effect on acute thermal responses. $n =$ number of flies. In all the plots, dark lines indicate mean and shaded regions indicate s.e.m. Black dashed lines in the plots represent temperature.



Extended Data Figure 10 | Activity profiles of flies in which clocks were sped up or slowed down in clock neurons. **a**, **b**, Sleep plots and sleep slope plots for two *Canton S* flies (**a**) and two *per^{01}* flies (**b**). We identified ZT'A' (red dashed line) as the local maximum slope value closest to ZT12 (blue dashed line, the onset of cooling) such that sleep slopes at ZT'A' and for the next 2 h were non-negative. The difference between these two values (12 − A), which is quantified as the anticipation index, is indicated in each of the plots. See Methods for more details. **c**, **d**, To quantify the extent to which the major activity peak anticipates the onset of cooling, we determined the phase of the daily peak of activity closest to ZT12 (the onset of cooling) for individual flies. Representative raw activity profiles of an individual wild-type *Canton S* fly (**c**) and *per^{01}* mutant fly (**d**). Red lines represent a smoothed activity profile created by means of a Butterworth low-pass filter (see Methods). The phases of the activity peak closest to the onset of cooling at ZT12 are indicated in the plots. ZT0 corresponds to the start of the heating phase and ZT12 to the start of the cooling phase. **e**–**l**, Averaged population activity plots over four days of temperature cycles for the genotypes indicated. The number of flies is shown in **m**. In all plots, dark lines indicate mean and shaded regions indicate s.e.m. Dashed lines in the plots represent temperature. **m**, The average phases of the major activity peak closest to the onset of the cooling phase. 0 marks the start of the cooling phase, which corresponds to ZT12 in the locomotor activity plots shown above. The average phase values were calculated by taking the population average of the phase of the major activity peaks of individual flies. Error bars represent s.e.m. Sample sizes are reported to the right of the plot. Negative average phase values indicate that the major activity peak occurred before the onset of the cooling phase, and positive values indicate that the major peak occurred after cooling onset. ANOVAs were conducted to compare *Clk4.1m>UAS-DBT(L)* flies to *Clk4.1m/+* and *UAS-DBT(L)/+* controls ($F_{2,95} = 26.1$, $P < 0.005$); *Clk4.1m>UAS-DBT(S)* flies to *Clk4.1m/+* and *UAS-DBT(S)/+* controls ($F_{2,76} = 6.5$, $P < 0.005$); *Clk856>UAS-DBT(L)* flies to *Clk856/+* and *UAS-DBT(L)/+* controls ($F_{2,78} = 18.4$, $P < 0.0005$); and *Clk856>UAS-DBT(S)* flies to *Clk856/+* and *UAS-DBT(S)/+* controls ($F_{2,77} = 25.1$, $P < 0.0005$); Tukey's tests are indicated. *$P < 0.05$, ***$P < 0.005$. Statistical analysis was performed using unpaired two-tailed Student's *t*-tests. **n**, Population averages of the phases of the activity peak closest to the onset of cooling (indicated by the zero line). The number of flies used in the analysis is reported in Extended Data Fig. 6a. Error bars represent s.e.m. An ANOVA was performed for each group of genotypes and Tukey's honest significant difference tests were conducted for pairwise comparisons. Statistical analysis was performed using unpaired two-tailed Student's *t*-tests, **$P < 0.005$. Individual *P* values are reported in the Source Data for this figure.

doi:10.1038/nature25744

Mechanical regulation of stem-cell differentiation by the stretch-activated Piezo channel

Li He[1], Guangwei Si[2], Jiuhong Huang[3], Aravinthan D. T. Samuel[2] & Norbert Perrimon[1,4]

Somatic stem cells constantly adjust their self-renewal and lineage commitment by integrating various environmental cues to maintain tissue homeostasis. Although numerous chemical and biological signals have been identified that regulate stem-cell behaviour, whether stem cells can directly sense mechanical signals *in vivo* remains unclear[1]. Here we show that mechanical stress regulates stem-cell differentiation in the adult *Drosophila* midgut through the stretch-activated ion channel Piezo. We find that *Piezo* is specifically expressed in previously unidentified enteroendocrine precursor cells, which have reduced proliferation ability and are destined to become enteroendocrine cells. Loss of *Piezo* activity reduces the generation of enteroendocrine cells in the adult midgut. In addition, ectopic expression of *Piezo* in all stem cells triggers both cell proliferation and enteroendocrine cell differentiation. Both the *Piezo* mutant and overexpression phenotypes can be rescued by manipulation of cytosolic Ca^{2+} levels, and increases in cytosolic Ca^{2+} resemble the Piezo overexpression phenotype, suggesting that Piezo functions through Ca^{2+} signalling. Further studies suggest that Ca^{2+} signalling promotes stem-cell proliferation and differentiation through separate pathways. Finally, *Piezo* is required for both mechanical activation of stem cells in a gut expansion assay and the increase of cytosolic Ca^{2+} in response to direct mechanical stimulus in a gut compression assay. Thus, our study demonstrates the existence of a specific group of stem cells in the fly midgut that can directly sense mechanical signals through Piezo.

Drosophila midgut stem cells have emerged as an attractive *in vivo* model for understanding adult stem-cell behaviours[2–4]. Like their mammalian counterparts, fly intestinal stem cells (ISCs) produce two major classes of cells that compose the adult intestinal epithelium: absorptive enterocytes and secretory enteroendocrine cells (EEs)[4]. Many extrinsic signals, including chemicals, nutrition, pathogens and cytokines, have been shown to regulate ISC proliferation and differentiation[4,5]. However, whether midgut stem cells can sense biomechanical signal remains unknown.

From a screen of GAL4 lines with expression in *Drosophila* midgut, we identified *PiezoP-GAL4* (BL59266, Bloomington *Drosophila* Stock Center)[6], a GAL4 transgene under control of a cloned enhancer/promoter region of *Piezo*, which was expressed in a subpopulation of Escargot (Esg)-positive stem cells in the adult fly midgut (Extended Data Fig. 1a). Piezo is a cation ion channel that directly senses mechanical tension in lipid bilayers[7]. It was initially identified in mammalian cells as a touching sensor[8], and was further found to be responsible for mechanoreception in different cell types[9]. The *Drosophila* genome encodes a single *Piezo* homologue, which has been characterized previously as a receptor for mechanotransduction in sensory neurons[6,10].

To represent the expression pattern of *Piezo* accurately, we directly knocked-in the GAL4 transgene into the *Piezo* locus after the first start codon through homologous recombination (referred to as *Piezo-GAL4* hereafter; Extended Data Fig. 1b). The expression of red fluorescent protein (RFP) driven by *Piezo-GAL4* showed a pattern similar to BL59266 in Esg^+ cells, but was also detected in some enterocytes located in the cardia and copper and iron regions (Fig. 1a, Extended Data Fig. 1c–f, h), which is consistent with published *Piezo* mRNA profiles along the midgut[11] (Extended Data Fig. 1g). Because Esg is expressed in both ISCs and enteroblasts (progeny of ISCs that are destined to become enterocytes), we used the ISC-specific marker *Delta-lacZ* and the enteroblast marker *Su*(*H*)*Gbe-lacZ* to identify $Piezo^+$ cells precisely. Notably, *Piezo* is expressed in a subpopulation (approximately 40%) of Delta-positive (Dl^+) cells, and is absent from enteroblasts (Fig. 1a, Extended Data Fig. 1i). We also noticed that almost all 'newborn' EEs (positive for both Esg and the EE-specific marker Prospero) are also $Piezo^+$, suggesting that $Piezo^+$ cells may represent enteroendocrine cell precursors (Fig. 1c, Extended Data Fig. 1k, l). Indeed, G-TRACE[12]-labelled progenies of $Piezo^+$ cells are primarily EEs (approximately 90%), rather than 11% EEs from Dl^+ ISCs, and 99% enterocytes from $Su(H)Gbe^+$ enteroblasts (Fig. 1d, e, Extended Data Fig. 1m–o). In addition, damage caused by bleomycin[13] or inhibition of Notch by the γ-secretase inhibitor DAPT[14] promotes the generation of both EEs and $Piezo^+$ cells (Fig. 1f, Extended Data Fig. 2a). Finally, ablation of $Piezo^+$ cells using the pro-apoptotic protein Reaper (Rpr) notably reduced not only the number of $Piezo^+$ cells but also enteroendocrine cell numbers after 4 weeks (Fig. 1g, h), and both cell types are recovered after one week of suppression of Rpr expression (Fig. 1g, h), suggesting that $Piezo^+$ cells are an important source of enteroendocrine cell generation. We further investigated whether $Piezo^+$ cells are self-regenerative or primarily derived from ISCs. First, mitotic $Piezo^+$ cells (marked by anti-phospho-histone3 (pH3) staining) represent only a small portion (~10%) of the total mitotic cells (Fig. 1i, Extended Data Fig. 2c–f), suggesting that $Piezo^+$ cells have reduced proliferation abilities compared to $Piezo^-$ stem cells. Bleomycin damage promotes the mitosis of both $Piezo^+$ and $Piezo^-$ cells without increasing the percentage of $Piezo^+$ mitotic cells, suggesting that an intrinsic mechanism limits the proliferation ability of $Piezo^+$ cells (Extended Data Fig. 2d, e). Finally, random green fluorescent protein (GFP)-marked clones generated from ISCs contain $Piezo^+$ cells, supporting the hypothesis that $Piezo^+$ cells are generated from ISCs (Extended Data Fig. 2g).

Taken together, our data suggest that previously considered Dl^+ ISCs are heterogeneous and composed of approximately 60% mitotic active multipotent ISCs ($Piezo^-$) and 40% less-mitotic unipotent $Piezo^+$ cells that mainly generate enteroendocrine cells. To avoid confusion with true ISCs (mitotic active and multipotent) and enteroblasts (occasionally referred to as Notch-active enterocyte progenitors), we refer to these $Piezo^+$ population as enteroendocrine precursors.

To investigate the function of Piezo, we analysed the phenotype of $Piezo^{KO}$, a null allele with a complete deletion of the *Piezo* coding sequence[6]. Midguts from $Piezo^{KO}$ homozygous flies showed no obvious phenotypes as compared to control flies during the early developmental and young adult stages, although Piezo is expressed in some stem cells during the larval and pupal stages (Extended Data Fig. 3). In wild-type

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Department of Physics, Center for Brain Science, Harvard University, Cambridge, Massachusetts 02142, USA. [3]School of Life Science and Technology, Tongji University, Shanghai 200092, China. [4]Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA.

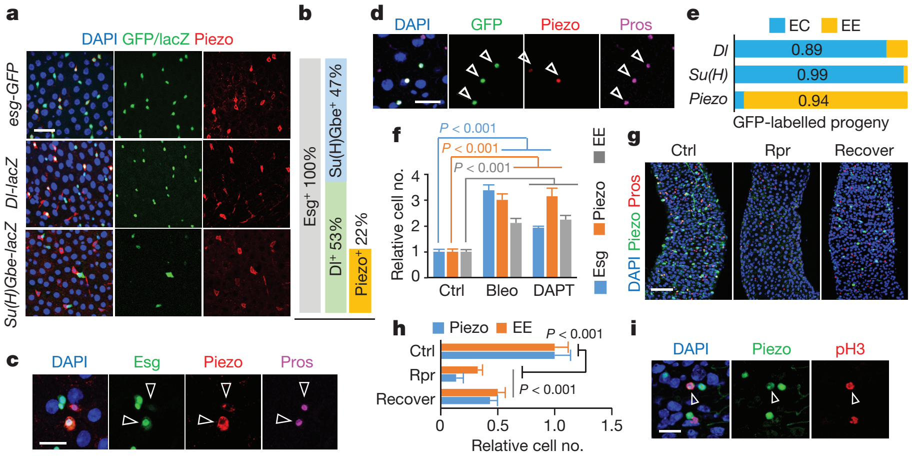


Figure 1 | Piezo$^+$ cells are EE precursors in the fly midgut. a, Piezo$^+$ cells (RFP$^+$, red) are Esg$^+$ (GFP, green) and Dl$^+$ (LacZ, green), but Su(H)Gbe$^-$ (LacZ, green). **b**, Percentage of Piezo$^+$ cells in Esg$^+$ cells. For Dl$^+$: $n=238$ (Dl$^+$), $n=457$ (Esg$^+$). For Piezo$^+$: $n=151$ (Piezo$^+$), $n=682$ (Esg$^+$). **c**, Newborn EEs (arrowheads) are Piezo$^+$ (mCherry$^+$, red; measured by *Piezo-GAL4*, *UAS-mCherry*) and Esg$^+$ (GFP$^+$, green; labelled by *esg-GFP*). **d**, Piezo$^+$ cells (RFP$^+$, red) generate GFP$^+$ EEs (arrowheads). **e**, Statistics of GFP$^+$ enterocytes (ECs) and EEs using *GAL4* lines *Piezo-GAL4*, *Su(H)Gbe-GAL4* and *Dl-GAL4*. Number of cells analysed: $n=561$ (*Dl*), $n=432$ (*Su(H)Gbe*), $n=90$ (*Piezo*). **f**, Treatment with bleomycin (bleo) and DAPT increases the number of Esg$^+$, Piezo$^+$ and EE cells. Areas quantified: $n=23$ (control; ctrl), $n=21$ (bleo), $n=32$ (DAPT). **g**, **h**, Elimination of Piezo$^+$ cells by conditional expression of Rpr. Areas quantified: $n=27$ (ctrl), $n=29$ (Rpr), $n=27$ (recover). **i**, pH3 staining (red) of mitotic enteroendocrine precursors (arrowheads). Data are mean and s.e.m. *P* values are from a two-tailed *t*-test. Scale bars, 20 μm (**a**), 10 μm (**c**, **d**, **i**) and 50 μm (**g**).

flies, the number of Esg$^+$ cells and EEs increases as the flies age[15]. However, in *Piezo^{KO}* mutants, the number of enteroendocrine cells, but not Esg$^+$ cells, does not increase (Fig. 2a, b), suggesting that the generation of EEs after adulthood is affected. Furthermore, *Piezo*-mutant clones generate 80% fewer EEs than controls, and this can be rescued by expressing GFP-tagged full-length Piezo (Fig. 2c, d). These data suggest that the reduced generation of EEs is an autonomous defect.

Previous studies have shown that *Piezo* functions through increases in cytosolic Ca^{2+} concentrations[16–19]. Consistently, knocking down *Stromal interaction molecule* (*Stim*), previously used as an effective target to decrease cytosolic Ca^{2+} levels[20], also led to the production of fewer EEs (Fig. 2c, d). Furthermore, increasing cytosolic Ca^{2+} by knocking down *plasma membrane calcium ATPase* (*PMCA*) or *Sarco/endoplasmic reticulum calcium ATPase* (*SERCA*) rescued and even reversed the reduction of EEs in the *Piezo* mutant (Fig. 2c, d). Overexpression of *Piezo* in Esg$^+$ cells caused an increase in both Esg$^+$ cells and EEs, and this phenocopied the increase of Ca^{2+} achieved by *SERCA* reduction, overexpression of *inositol-1,4,5-trisphosphate receptor* (*InsP3R*, also known as *Itp-r83A*), *Stim* and *Orai* (*olf186-F*), and *PMCA* knockdown (Fig. 2e, Extended Data Figs 4a–c, 5). Calcium imaging shows that cytosolic Ca^{2+} is significantly increased by *Piezo* overexpression in the stem cells (Extended Data Fig. 6a–d, Supplementary Videos 1, 2). The *Piezo*-overexpression phenotype is suppressed by reducing cytosolic Ca^{2+} using RNA interference (RNAi) that targets either *Stim* or *InsP3R* (Fig. 2e, Extended Data Fig. 4a–c). Finally, damage caused by bleomycin triggers an upregulation of Ca^{2+} and an increase in the number of Esg$^+$ and EE cells in both wild-type and *Piezo^{KO}* midguts, supporting the idea that Ca^{2+} is the downstream effector of Piezo (Extended Data Figs 5d, e, 6e–h).

Inhibition of Notch signalling has been shown to promote both ISCs renewal and EE differentiation[14,21], even in EE progenitors that already

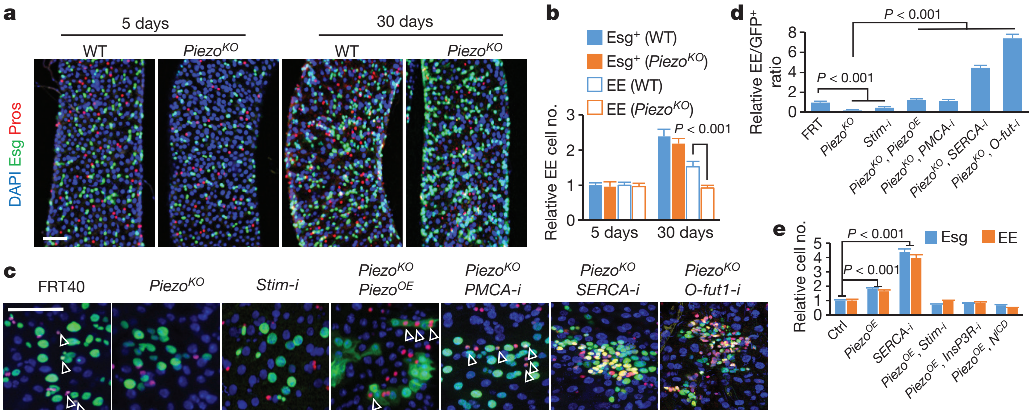


Figure 2 | Piezo regulates EE differentiation through cytosolic Ca^{2+}. a, **b**, Midgut of flies homozygous for the null mutant *Piezo^{KO}* shows reduced EE generation 30 days after eclosion. Areas quantified: $n=32$ (wild type (WT) 5 days), $n=32$ (WT 30 days), $n=35$ (*Piezo^{KO}* 5 days), $n=32$ (*Piezo^{KO}* 30 days). **c**, **d**, MARCM clones of cells with indicated genotypes (arrowheads indicate GFP$^+$ EEs). The ratio of EEs in the clone (normalized to control) is quantified. The '-*i*' suffix denotes RNAi. Number of clones quantified: $n=32$ (*FRT*), $n=35$ (*Piezo^{KO}*), $n=26$ (*Stim-i*), $n=28$ (*Piezo^{KO}*, *Piezo^{OE}*), $n=31$ (*Piezo^{KO}*, *PMCA-i*), $n=35$ (*Piezo^{KO}*, *SERCA-i*), $n=28$ (*Piezo^{KO}*, *O-fut1-i*). **e**, Esg$^+$ and EE cell numbers were quantified in the indicated midgut-expressing genes using *esg-GAL4*. Number of areas quantified: $n=22$ (ctrl), $n=28$ (*Piezo^{OE}*), $n=23$ (*SERCA-i*), $n=21$ (*Piezo^{OE}*, *Stim-i*), $n=24$ (*Piezo^{OE}*, *InsP3R-i*), $n=26$ (*Piezo^{OE}*, *N^{ICD}*). Data are mean and s.e.m. *P* values are from a two-tailed Student's *t*-test. Scale bars, 50 μm (**a**) and 25 μm (**c**).

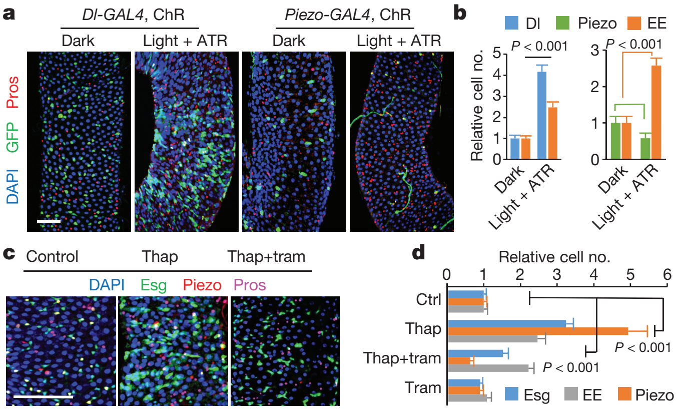


Figure 3 | Cytosolic Ca^{2+} triggers cell proliferation and EE differentiation through different mechanisms. a, **b**, Increase in cytosolic Ca^{2+} by channelrhodopsin (ChR) in Dl^{+} and Piezo^{+} enteroendocrine precursors. Dl^{+}, Piezo^{+} and EE cell numbers are quantified (**b**). Number of areas quantified: $n = 28$ (dark, *Dl-GAL4*), $n = 30$ (light + ATR, *Dl-GAL4*), $n = 30$ (dark, *Piezo-GAL4*), $n = 31$ (light + ATR, *Piezo-GAL4*). **c**, **d**, Midguts of flies treated with thapsigargin (thap) and trametinib (tram). Number of areas quantified: $n = 29$ (ctrl), $n = 31$ (thap), $n = 32$ (thap + tram), $n = 29$ (tram). Data are mean and s.e.m. *P* values are from a two-tailed *t*-test. Scale bars, 50 μm.

have low Notch activity[22]. In addition, increases in cytosolic Ca^{2+} have been found to inhibit Notch activity in both cultured mammalian cells and flies[23,24]. We therefore tested whether Piezo functions through Notch inhibition by increasing cytosolic Ca^{2+}. Indeed, blocking Notch activation by knocking down the Notch-processing enzyme *O*-fucosyltransferase (encoded by *O-fut1*) reverses the Piezo-knockout phenotype (Fig. 2c, d, Extended Data Fig. 4h), and increasing Notch activity by expression of the Notch intracellular domain (N^{ICD}) blocks the phenotypes of both *Piezo* overexpression and *SERCA* knockdown (Fig. 2e, Extended Data Fig. 4a–g). Furthermore, overexpressing *Piezo* in Esg^{+} cells produced more Dl^{+} stem cells, consistent with a reduction in Notch activity (Extended Data Fig. 6i, j). Finally, neither *Piezo* overexpression nor *SERCA* knockdown had any effect in enteroblast cells (in which Notch has already been activated), suggesting that Notch signalling is the primary target (Extended Data Fig. 6k, l). Taken together, our data suggest that *Piezo* promotes EE differentiation by increasing cytosolic Ca^{2+} and inhibition of Notch.

To investigate the function of Ca^{2+} further, we used channelrhodopsin (ChR) to increase cytosolic Ca^{2+} levels optogenetically. Activation of ChR in Dl^{+} cells promotes both ISC proliferation and EE production, resembling the *Piezo*-overexpression phenotype (Fig. 3a, b, Extended Data Fig. 7a–d). This ChR-induced phenotype is blocked by knockdown of both *Stim* and *InsP3R*, suggesting that the effect is Ca^{2+} dependent (Extended Data Fig. 7e, f). In addition, activation of ChR in Piezo^{+} enteroendocrine precursors significantly increased the number of EE cells at the expense of precursor cells, suggesting an increase in the differentiation of enteroendocrine precursor to EEs (Fig. 3a, b). A recent study showed that Piezo activation promotes cell proliferation through Ca^{2+}-induced phosphorylation of ERK[19]. Consistently, overexpression of *Piezo* in Esg^{+} cells increases phospho-ERK staining (Extended Data Fig. 7g). However, reducing ERK signalling through *Ras* knockdown or blocking cell proliferation by *yorkie* (*yki*) RNAi only affects cell proliferation, and not EE differentiation, in *Piezo*-overexpressing cells (Extended Data Fig. 7h–k), suggesting that Piezo promotes EE differentiation independently of proliferation. Consistently, increasing cytosolic Ca^{2+} using the SERCA inhibitor thapsigargin significantly increased stem-cell proliferation and EE generation (Fig. 3c, d). Further blocking mitosis using the MEK inhibitor trametinib only reduced thapsigargin-trigged proliferation, but not the increase in EE differentiation (Fig. 3c, d, Extended Data Fig. 7l–n). Ca^{2+} imaging showed that Ca^{2+} is increased in stem cells treated by thapsigargin, which is not blocked by trametinib (Extended Data Fig. 7o–q, Supplementary Video 3). Taken together, these data suggest that increases in cytosolic Ca^{2+} promote cell proliferation (through ERK phosphorylation) and cell differentiation (though Notch inhibition) in a cell-context-dependent manner.

To test whether mechanical challenges from food digestion can activate Piezo, we increased the mechanical load in the gastrointestinal

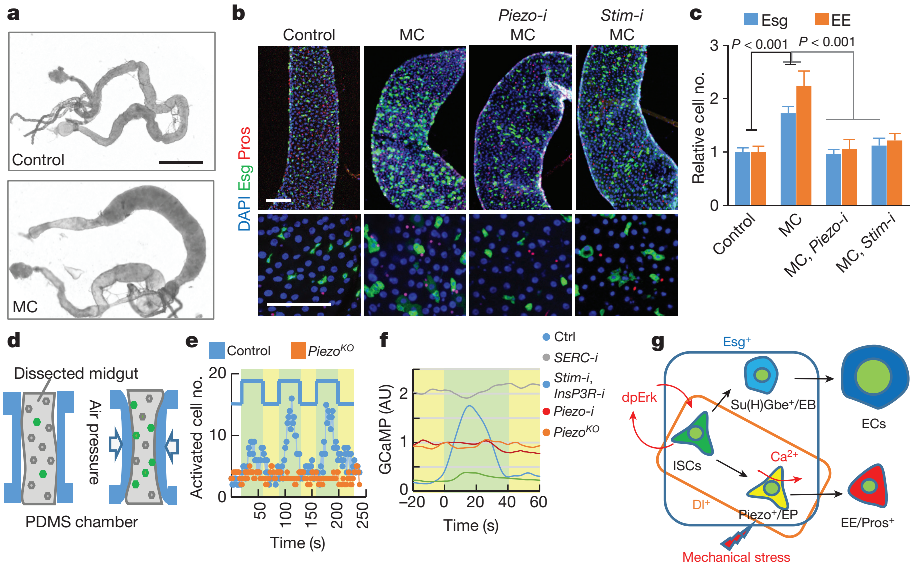


Figure 4 | Mechanical stress increases cytosolic Ca^{2+} through Piezo. a, Midgut of fly fed on food containing methylcellulose (MC). **b**, **c**, Methylcellulose feeding increases Esg^{+} (GFP^{+}, green) and EE cell numbers in the midguts, which is blocked by *Piezo-i* and *Stim-i*. Number of areas quantified: $n = 25$ (ctrl), $n = 23$ (MC), $n = 20$ (MC + *Piezo-i*), $n = 25$ (MC + *Stim-i*). **d**, An illustrated microfluidic channel that holds and compresses the midgut for *ex vivo* mechanical trigger experiments. **e**, Representative example of three cycles of consecutive mechanical activation. Number of Ca^{2+} cells is plotted over time. Green, compression period; yellow, relaxation period. **f**, Average GCaMP activity during compression from control, *Piezo^{KO}*, *Piezo-i*, *SERCA-i* and *Stim-i* + *InsP3R-i* flies. **g**, Model for mechanical regulation of enteroendocrine precursor differentiation in the fly midgut. Ca^{2+} has different roles in ISCs (proliferation) and enteroendocrine precursors (differentiation). dpErk, extracellular signal-regulated kinase (also known as Rl). Data are mean and s.e.m. *P* values are from a two-tailed Student's *t*-test. Scale bars, 10 mm (**a**) and 50 μm (**b**).

tract by feeding flies with food containing the indigestible fibre methylcellulose, which is a widely used food thickener and ingredient for cell culture. This methylcellulose food induces an 'over-full' phenotype, as fly midguts from approximately 10–15% of flies after 4–5 days of methylcellulose feeding showed a significant increase in diameter (Fig. 4a, Extended Data Fig. 8). Midguts with increased diameter showed a significant increase in the number of Esg$^+$ cells and EEs (Fig. 4b, c), as well as Piezo$^+$ enteroendocrine precursor cells (Extended Data Fig. 8g–j). This effect is blocked by either *Piezo* knockdown or the null mutant (*Piezo* RNAi or *Piezo^{KO}*, respectively) (Fig. 3b, c, Extended Data Fig. 8k, l). Live-cell imaging of Ca^{2+} activities shows an increase of average Ca^{2+} levels in methylcellulose-fed flies, suggesting that the phenotype is related to increased Ca^{2+} levels (Extended Data Fig. 8n–q, Supplementary Video 4). Indeed, this over-full phenotype is blocked by reducing cytosolic Ca^{2+} (Fig. 4b, c, Extended Data Fig. 8n–q, Supplementary Video 4), suggesting that the mechanical stress generated by the indigestible food promotes EE generation through Piezo activation and subsequent increases in cytosolic Ca^{2+}. As Piezo is mainly enriched in enteroendocrine precursor cells, the increase of stem-cell proliferation may be caused by either a feedback signal from the increased EE generation[25] or low levels of Piezo present in the ISCs.

To test directly whether mechanical forces can activate enteroendocrine precursor cells, we engineered a microfluidic chip that can hold a dissected fly midgut and generate a mechanical compression through controlled air pressure (Fig. 4d, Extended Data Fig. 9a–d). Using this device, we recorded the calcium signal in Dl$^+$ stem cells of the fly midguts (*Piezo-GAL4* was tested initially but was not used owing to the low GCAMP6s expression). Significantly more stem cells showed high cytosolic Ca^{2+} upon mechanical compression, and this activation was triggered transiently by the change in tissue shape, as Ca^{2+} activity returned to normal within approximately 20 s, even in the presence of constant compression (Fig. 4e, Supplementary Video 6). This mechanically triggered Ca^{2+} activity is significantly reduced in either *Piezo^{KO}* or *Piezo* RNAi midguts (Fig. 4e, Extended Data Fig. 9e–g, Supplementary Videos 7, 8). Finally, either increases of cytosolic Ca^{2+} through *SERCA* knockdown or decreases of cytosolic Ca^{2+} through *Stim* and *InsP3R* knockdown render the cells irresponsive to the mechanical stimulus (Fig. 4f, Extended Data Fig. 9h–l, Supplementary Videos 9, 10). These data suggest that Ca^{2+} levels in Piezo$^+$ cells can be regulated by a transient mechanical stimulus, which may be generated by repeated vascular muscle contractions during digestion.

In conclusion, we have demonstrated that a new population of unipotent stem cells (enteroendocrine precursors) can directly sense mechanical signals *in vivo* to adjust their differentiation accordingly, and that this mechanosensing is mediated through Piezo activation and cytosolic Ca^{2+} increase. Our findings suggest a potential direct linkage between food digestion with generation of EEs, which regulate various physiological functions, including stem-cell proliferation, intestinal motility, digestion and appetite[25,26]. Such a mechanism may enable the midgut to respond to particular mechanical challenges and maintain tissue homeostasis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 September 2016; accepted 12 January 2018.
Published online 7 February 2018.

1. Vining, K. H. & Mooney, D. J. Mechanical forces direct stem cell behaviour in development and regeneration. *Nat. Rev. Mol. Cell Biol.* **18,** 728–742 (2017).
2. Micchelli, C. A. & Perrimon, N. Evidence that stem cells reside in the adult *Drosophila* midgut epithelium. *Nature* **439,** 475–479 (2006).
3. Ohlstein, B. & Spradling, A. The adult *Drosophila* posterior midgut is maintained by pluripotent stem cells. *Nature* **439,** 470–474 (2006).
4. Li, H. & Jasper, H. Gastrointestinal stem cells in health and disease: from flies to humans. *Dis. Model. Mech.* **9,** 487–499 (2016).
5. Lemaitre, B. & Miguel-Aliaga, I. The digestive tract of *Drosophila melanogaster*. *Annu. Rev. Genet.* **47,** 377–404 (2013).
6. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. The role of *Drosophila* Piezo in mechanical nociception. *Nature* **483,** 209–212 (2012).
7. Coste, B. *et al.* Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature* **483,** 176–181 (2012).
8. Coste, B. *et al.* Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330,** 55–60 (2010).
9. Volkers, L., Mechioukhi, Y. & Coste, B. Piezo channels: from structure to function. *Pflugers Arch.* **467,** 95–99 (2015).
10. Suslak, T. J. *et al.* Piezo is essential for amiloride-sensitive stretch-activated mechanotransduction in larval *Drosophila* dorsal bipolar dendritic sensory neurons. *PLoS ONE* **10,** e0130969 (2015).
11. Buchon, N. *et al.* Morphological and molecular characterization of adult midgut compartmentalization in *Drosophila*. *Cell Reports* **3,** 1725–1738 (2013).
12. Evans, C. J. *et al.* G-TRACE: rapid GAL4-based cell lineage analysis in *Drosophila*. *Nat. Methods* **6,** 603–605 (2009).
13. Amcheslavsky, A., Jiang, J. & Ip, Y. T. Tissue damage-induced intestinal stem cell division in *Drosophila*. *Cell Stem Cell* **4,** 49–61 (2009).
14. Ohlstein, B. & Spradling, A. Multipotent *Drosophila* intestinal stem cells specify daughter cell fates by differential notch signaling. *Science* **315,** 988–992 (2007).
15. Choi, N. H., Kim, J. G., Yang, D. J., Kim, Y. S. & Yoo, M. A. Age-related changes in *Drosophila* midgut are associated with PVF2, a PDGF/VEGF-like growth factor. *Aging Cell* **7,** 318–334 (2008).
16. Cinar, E. *et al.* Piezo1 regulates mechanotransductive release of ATP from human RBCs. *Proc. Natl Acad. Sci. USA* **112,** 11783–11788 (2015).
17. Pathak, M. M. *et al.* Stretch-activated ion channel Piezo1 directs lineage choice in human neural stem cells. *Proc. Natl Acad. Sci. USA* **111,** 16148–16153 (2014).
18. Li, J. *et al.* Piezo1 integration of vascular architecture with physiological force. *Nature* **515,** 279–282 (2014).
19. Gudipaty, S. A. *et al.* Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature* **543,** 118–121 (2017).
20. Deng, H., Gerencser, A. A. & Jasper, H. Signal integration by Ca^{2+} regulates intestinal stem-cell activity. *Nature* **528,** 212–217 (2015).
21. Guo, Z. & Ohlstein, B. Stem cell regulation. Bidirectional Notch signaling regulates *Drosophila* intestinal stem cell multipotency. *Science* **350,** aab0988 (2015).
22. Sallé, J. *et al.* Intrinsic regulation of enteroendocrine fate by Numb. *EMBO J.* **36,** 1928–1945 (2017).
23. De Ford, C. *et al.* The clerodane diterpene casearin J induces apoptosis of T-ALL cells through SERCA inhibition, oxidative stress, and interference with Notch1 signaling. *Cell Death Dis.* **7,** e2070 (2016).
24. Roti, G. *et al.* Complementary genomic screens identify SERCA as a therapeutic target in NOTCH1 mutated cancer. *Cancer Cell* **23,** 390–405 (2013).
25. Amcheslavsky, A. *et al.* Enteroendocrine cells support intestinal stem-cell-mediated homeostasis in *Drosophila*. *Cell Reports* **9,** 32–39 (2014).
26. Harrison, E., Lal, S. & McLaughlin, J. T. Enteroendocrine cells in gastrointestinal pathophysiology. *Curr. Opin. Pharmacol.* **13,** 941–945 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Binari, W. Song and C. Villalta for technical support; C. Xu, S. Mohr and D. Doupe for comments on the manuscript; and G. Hasan for sharing reagents. This work was supported by the Damon Runyon Cancer Research Foundation (L.H.) and a grant from the National Institutes of Health (R21DA039582). N.P. is an investigator of the Howard Hughes Medical Institute. G.S. and A.D.T.S. are supported by the National Institutes of Health (P01GM103770).

Author Contributions J.H. and L.H. performed the initial GAL4 expression screen in fly gut. L.H. and N.P. designed the experiments. L.H. performed the Piezo-related experiments and analysed the data. G.S. and A.D.T.S. designed and fabricated the microfluidic chip and together with L.H. optimized the experimental conditions. L.H. and N.P. wrote the manuscript with input from all of the authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to L.H. (lihe@genetics.med.harvard.edu) or N.P. (perrimon@rascal.med.harvard.edu).

Reviewer Information *Nature* thanks L. O'Brien, K. Venkatachalam and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

***Drosophila* stocks and culture.** The following strains were obtained from the Bloomington *Drosophila* Stock Center: *UAS-mtdTomato3XHA* (BL30124), *UAS-tdTomato* (BL3321, BL3322), *UAS-IVS-NES-jRGECO* (BL63795), *UAS-IVS-GCaMP6s* (BL42746), *UAS-mCherry.CAAX* (BL59021), *UAS-mCherry.nls* (BL38424), *UAS-CsChrimson* (BL55134), *tub-GAL80*^{TS} (BL7016), *UASp-Act5C-mRFP* (BL24778), *UAS-mCD8-GFP* (BL32185), *hsFLP*; *Sco/CyO* (BL1929), *UAS-Stim* (BL41757), *UAS-InsP3R* (BL30742), *PiezoP-GAL4* (with cloned promoter, BL59266), *UAS-O-fut1-i* (BL9377), *UAS-Notch*^{ICD} (BL52008), *UAS-Rpr* (BL5823), *Piezo*^{KO} (BL58770), *UAS-Piezo-GFP/CyO* (BL58772), *UAS-Piezo-GFP/TM6B* (BL58773)[6], *UAS-ttk69-i* (BL26315, BL36748), *Act-(FRT.Stop)lacZ*, *Ubi-(FRT.Stop)Stinger/CyO* (isolated from BL51308), and G-Trace fly: *UAS-RedStinger*, *UAS-Flp1.D*; *Ubi-(FRT.Stop)Stinger/CyO* (BL28280). RNAi lines were as previously reported[20]: *UAS-SERCA-i* (BL25928), *UAS-Stim-i* (BL27263, BL52911), *UAS-InsP3R-i* (BL25937, BL51686), and *UAS-PMCA-i* (BL31572). *UAS-Ras1-i* (106642), *UAS-Yki-i* (104523), *UAS-Piezo-i* (2796), *UAS-ase-i* (108511) and *UAS-ttk69-i* (101980) as previously reported[10,27] were from the Vienna *Drosophila* RNAi Center. *esg-GFP* was from D. Doupe. *Su(H)Gbe-lacZ* was from P. Saavedra. The fly stock for mosaic analysis, *hsFLP*, *tub-GAL4*, *UAS-nlsGFP*; *FRT40*, *tub-GAL80*, was from K. Kim; *Su(H)Gbe-GAL4* and *Dl-GAL4* were from S. X. Hou[28], *UAS-Orai* was from G. Hasan, *esg-GAL4*, *UAS-nlsGFP* and *Dl-lacZ* were from laboratory stocks. Flies were reared on standard cornmeal/agar medium supplemented with yeast. Adult flies were entrained in 12 h:12 h light–dark cycles at 25 °C unless specifically stated otherwise.

To prepare methylcellulose food, 10% (w/w) methylcellulose (Sigma-Aldrich, 274429) was added to 5% sucrose solution and stirred until fully dissolved. Adult flies 5–7 days after hatching were water-starved (soaked filter paper) for 1 day at 29 °C, and transferred to vials with methylcellulose or control food (5% sucrose-soaked filter paper). Food was changed every other day. Fly midguts with a considerably enlarged diameter (>50% increase compared with the normal section of the same midgut) were counted as enlarged methylcellulose-fed gut (∼10–15% of total dissected midguts).

DAPT (4 μM; Sigma-Aldrich, D5942), bleomycin (10 μg ml^{-1}; Calbiochem, 203408), thapsigargin (0.5 μM; Tocris, 1138) and trametinib (10 μM; Selleckchem, S2673) were used for chemical treatment. All feeding experiments were performed using 5% sucrose-saturated filter paper unless stated otherwise.

For the lineage-tracing experiments[12], the temperature-sensitive GAL80^{TS} transgene was used to suppress the early activity of GAL4 before adulthood. Flies (4–5 days old) were incubated at 32 °C for 1 day to activate GAL4 and then maintained at 25 °C for 7 days. For *Piezo-GAL4*, flies were incubated at 32 °C for 4 days and then maintained at 25 °C for 3 days because of its low activity. Lineage tracing of methylcellulose-fed flies was done by induction of flies for 4–5 days at 32 °C and when feeding the flies on 5% sucrose plus 10% methylcellulose for 4 days at 25 °C. To visualize the GAL4-expressing cells, flies were shifted to 32 °C overnight before analysis. To create random clones using *hsFLP*; *Ubi-(FRT.Stop)Stinger*, we heat-shocked the 3–4-day-old adult flies at 37 °C for 30 min, and then kept them at 25 °C for 2 weeks.

For the mosaic analysis with a repressible cell marker (MARCM) experiments[29], 4–5-day-old flies were heat-shocked three times at 37 °C for 1 h within 1 day. Then flies were maintained at 25 °C, except for the flies containing RNAi, which were maintained at 32 °C to increase the expression of the double-stranded RNAs. Temperature has no significant effect on the ratio of EEs in the progenies (data not shown). Midguts from female flies were analysed after 14 days. (GFP-positive clones were induced by transient incubation at 32 °C, then flies were kept at 25 °C for 10 days and 32 °C overnight before analysis.)

Immunofluorescence imaging. Immunostaining of *Drosophila* midguts was performed as previously described[30]. The following primary antibodies were used: mouse anti-Prospero (1:50; Developmental Studies Hybridoma Bank, MR1A), rabbit anti-phospho-histone H3 (1:1,000; Millipore, 06–570); mouse anti-haemagglutinin (Abcam, ab18181), rabbit anti-dpErk1/2 (1:500; Cell Signaling, 4370), mouse anti-Delta (1:50; Developmental Studies Hybridoma Bank, C594.9B), mouse anti-β-galactosidase (1:400; Promega, Z3781), rabbit anti-tachykinin (1:5,000: ref. 31). Secondary antibodies were goat anti-rabbit and anti-mouse IgGs conjugated to Alexa 555 and Alexa 647 (used at 1:500, Thermo Fisher Scientific, A-21428, A-21244, A-21235, A-21422). Fly guts were mounted in Vectashield with DAPI (Vector Laboratories). In all micrographs, blue staining shows the nuclear marker DAPI. Fluorescence micrographs were acquired with a Zeiss LSM 780 confocal microscope. All images were adjusted and assembled in NIH ImageJ.

CRISPR–Cas9 genome editing. Guide RNAs (gRNAs) targeting the start codon of *Piezo* were designed using the 'find CRISPRs' online tool (http://www.flyrnai.org/crispr2/)[32,33]. The genome-editing efficiency of different candidate gRNAs was tested in tissue culture using T7 endonuclease assay[34], and the following sequence with highest cutting efficiency was used: CTGGAGGAGAACGGCGCCGG.

Genomic fragments approximately 1 kb from the upstream and downstream of the start codon were amplified from fly genome using the following primers: upstream forward: 5′-CTTCGGTACCGGATCACTGTGCATGTGAGGCATTA-3′, upstream reverse: 5′-GCTTCATTTTGGATCACTCAGACTCCGACTCCAAC-3′; downstream forward: 5′-CGGCGGCCGCTCTAGTCAGCTATGCGTGCATGGT-3′, downstream reverse: 5′-AAGCTGGGTGTCTAGGGGAATGTGGTAGGCAAACTA-3′.

Genomic fragments were cloned upstream and downstream of GAL4-SV40 in pENTR vector by In-fusion (Clontech) to make the donor construct.

For CRISPR–Cas9-mediated homologous recombination, gRNA in pCFD3 (0.2 μg μl^{-1}) and donor DNA (0.5 μg μl^{-1}), were co-injected into the embryos of *nos-Cas9/attP2* flies[35]. Knock-in flies were selected by genomic PCR using following primers from insertion and *Piezo* gene: upstream forward: 5′-CCCACAATTTCGCACTCTTT-3′, upstream reverse: 5′-GTCTTCACGGGGAAAAATGA-3′; downstream forward: 5′-GTGGTTTGTCCAAACTCATCAATG-3′, downstream reverse 5′-CGGACAGCAGGAAAATGAGA-3′.

Piezo-GAL4 knock-in homozygous flies are viable and fertile. Quantitative PCR (qPCR) of whole adult flies showed that *Piezo* mRNA from homozygous *Piezo-GAL4* knock-in flies was reduced by ∼50% compared to *Piezo-GAL4/CyO*. The mRNA of *Piezo* from *Piezo-GAL4/CyO* was not significantly different from wild-type flies. Also, qPCR of *Piezo*^{KO} (BL58770) is consistent with this allele being a complete null[6] as it showed a more than 95% reduction of *Piezo* mRNA.

Optogenetic activation of CsChrimson in fly midgut. Red-shifted channelrhodopsin CsChrimson[36] was used to increase cytosolic Ca^{2+} in stem cells by light. *UAS-CsChrimson* was expressed using either *Dl-GAL4* or *Piezo-GAL4*. All crosses and the early development of flies were performed under dark conditions at 18 °C. The experiment was done at 25 °C. Adult flies were kept either on 2% agarose containing 5% sucrose and 1% yeast extract in the dark, or on 2% agarose containing 5% sucrose, 1% yeast extract and 50 mM all-trans-retinal (ATR) in the presence of orange–red light from LED. Two 1-metre SMD5050 RGB LED strips (total power ∼2 × 4 W, eTopxizu) was attached to the inner wall of a cylinder chamber (∼10 cm in diameter and 15 cm in height) covered by aluminium foil to enhance the light intensity (Extended Data Fig. 7a). The RGB LED strip was set at constant maximal brightness with green (500–560 nm) and red (600–650 nm) LED units on (estimated light intensity ∼ 2.5 mW cm^{-2}). The power of the LED is controlled manually to maintain 12 h/12 h on/off circadian rhythms. Flies were kept under the indicated condition for 2 weeks before analysis.

Calcium imaging. Cytosolic Ca^{2+} was monitored in ISCs using the red fluorescent indicator RGECO[37]. GFP was used as an internal control and an indicator of stem cells and enteroblasts. Young adult flies (4–5 days after eclosion) were first incubated at 32 °C for 5–7 days before the experiment. For live-cell imaging experiment, dissected intact midgut was cultured in adult-hemolymph-like (AHL) medium plus 2% fetal bovine serum (FBS). The addition of FBS into the AHL moderately increases the average cytosolic Ca^{2+} level and reduced the oscillation frequency, but allows a longer maintenance of dissected midgut under normal condition up to 5–6 h. Air-permeable lummox dish (SARSTEDT, 94.6077.331) was used as the imaging device as previously described[38]. Images of anterior midgut area were captured on Zeiss LSM 780 confocal microscope equipped with definite focus using Plan-Neofluar 25×/oil numerical aperture (NA) 0.8 lens. A *z*-stack of dual-colour images (488 nm excitation/500–550 nm detection for GFP, and 561 nm excitation/580–650 nm detection for RGECO) was recorded every 20 s. Both colour channels were recorded simultaneously with line-based scanning. Images were manually analysed in NIH ImageJ.

Microfluidic chip design and operation. The fly gut was immobilized and force stimuli were applied in a microfluidic chip. The design took advantage of the pressure sensitivity of the poly material (PDMS, the building materials of the microfluidics), and had been applied in previous studies of *Caenorhabditis elegans*[39]. The chip was designed using the software of Tanner L-Edit and fabricated following standard microfluidics fabrication procedures[40]. The layout of the design is shown in Extended Data Fig. 9. The middle channel was designed for loading and holding the gut, with a size of 6 mm long and 200 μm wide. The two side channels delivered the pressure, with a size of 1 mm long and 450 μm width. The membrane in between is 70 μm wide, and was used for squeezing the guts when pressures were applied. The pattern was transferred onto a silicon wafer via photoresist with the height of 200 μm, which was then transferred to PDMS and bonded with glass. To achieve the desired softness, the PDMS was mixed 20:1 with the cross-linker.

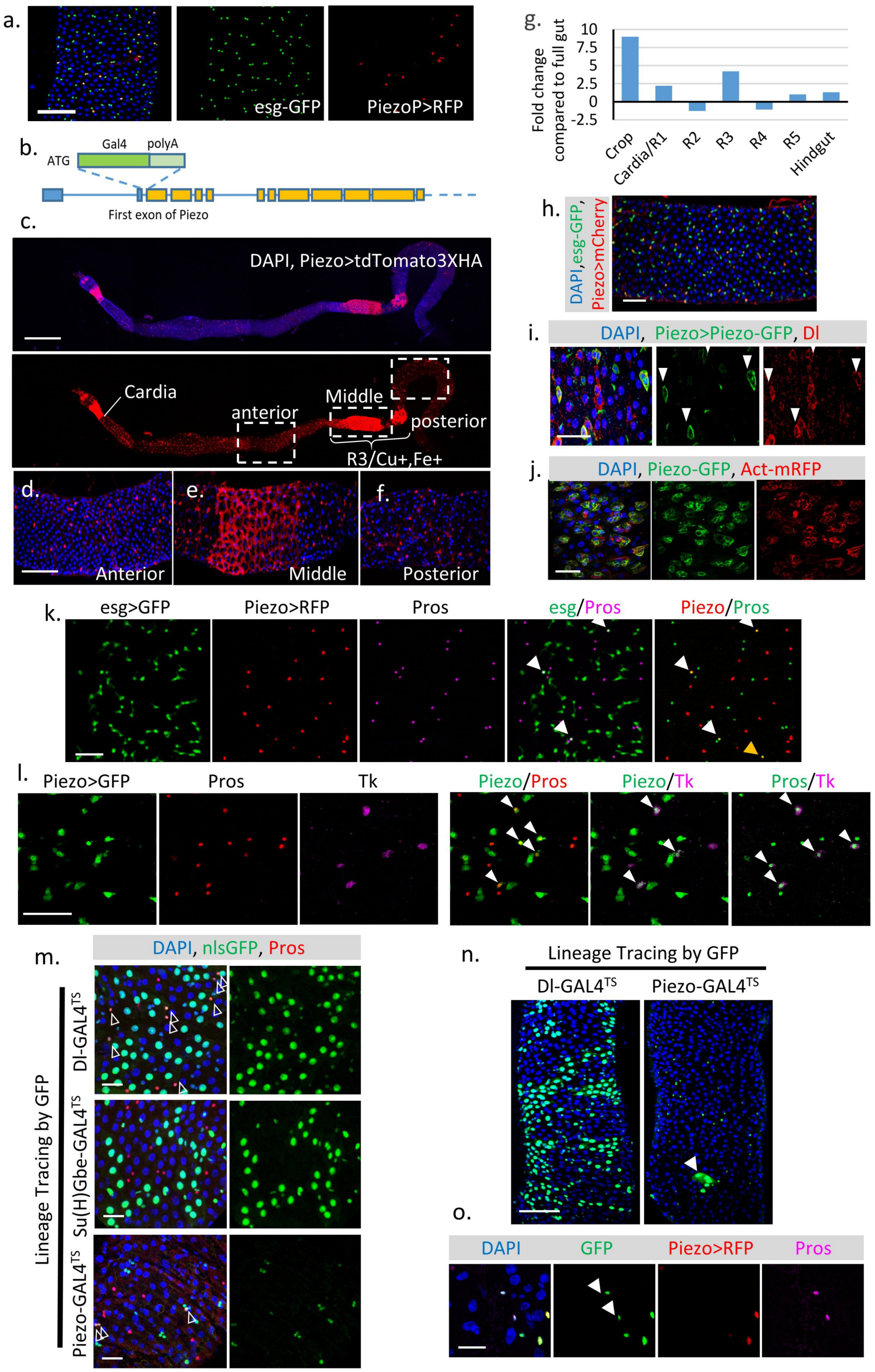
Freshly dissected fly midguts were loaded in the channel inlet with the anterior part of the gut located in the middle between the two membranes. In the device, compressed air is connected to the side channels via a bidirectional switch. In the off state, the side channels are at the atmospheric pressure, and no pressure is applied to the gut. When switched to the on state, compressed air presses the PDMS membrane and squeezes the gut. The ratio of the channel width reduction was ∼30% during the compression and the relaxation time of the PDMS membrane was $\sim$1 s. Ca^{2+} signals were indicated by GCAMP6s[41] and captured using a Zeiss LSM 780 confocal microscope equipped with a definite focus using Plan-Neofluar 10×/0.30 lens. The anterior midgut area was recorded as time-lapse of *z*-stacks capturing the whole depth of the midgut every 2 s. GCAMP6s emission was excited at 448 nm and recorded at 500–550 nm and tdTomato was excited at 561 nm and recorded at 580–610 nm. Ca^{2+} imaging experiments were done with identical acquisition parameters for consistency. Images from the experiment were projected using maximum intensity projection and analysed using a macro in ImageJ to automatically detect the number of GFP-positive cells in each frame. Tracing of Ca^{2+} signals in individual stem cells was done using the *Z*-axis profiling function of NIH ImageJ. The Ca^{2+} signal in individual stem cells during mechanical compression was tacked manually.

qPCR. Total RNA was extracted from 5–7-day-old female by TRIZOL reagent (Thermo Fisher Scientific), converted to cDNA template after DNase I treatment and purification using QIAGEN RNeasy kit. qPCR was performed using SYBR Green with *Gapdh1* and α*Tub84B* as internal controls. *Piezo* mRNA was detected by two pairs of independent primers (Supplementary Table 2).

Statistics and reproducibility. All the images presented and used for quantification are from the anterior region of adult female fly midgut for consistency. Two or three square areas (10,000 μm^2 unless specified otherwise) were randomly selected from each midgut and quantified automatically using the cell counting function of NIH ImageJ. All experiments were independently biologically repeated twice (unless specified otherwise) with similar results presented in the figures. No randomization or blinding was used. Statistical analysis was performed using Microsoft Excel. All *P* values were determined by two-tailed Student's *t*-test with unequal variances. Sample sizes were chosen empirically based on the observed effects and listed in the figure legends.

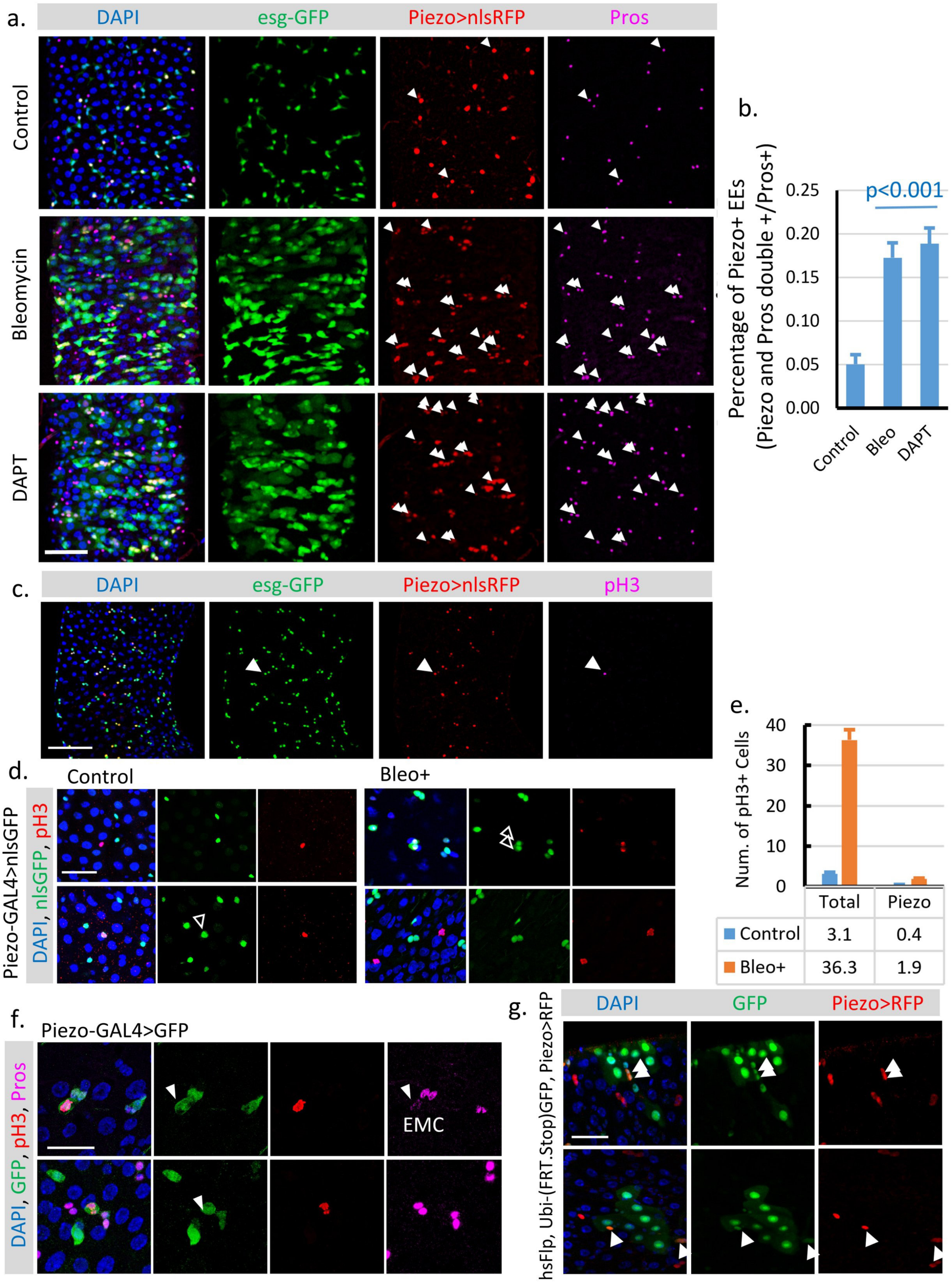
Data availability. All relevant data have been included in the paper and its Supplementary Information. Original quantifications of different cell numbers are listed in the Supplementary Data. Complete genotypes information is provided in Supplementary Table 1. Original data that support the findings of this study are available from the corresponding author upon request.

27. Xu, C., Luo, J., He, L., Montell, C. & Perrimon, N. Oxidative stress induces stem cell proliferation via TRPA1/RyR-mediated Ca^{2+} signaling in the *Drosophila* midgut. *eLife* **6,** e22441 (2017).
28. Zeng, X., Chauhan, C. & Hou, S. X. Characterization of midgut stem cell- and enteroblast-specific GAL4 lines in *Drosophila*. *Genesis* **48,** 607–611 (2010).
29. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker (MARCM) for *Drosophila* neural development. *Trends Neurosci.* **24,** 251–254 (2001).
30. Karpowicz, P., Perez, J. & Perrimon, N. The Hippo tumor suppressor pathway regulates intestinal stem cell regeneration. *Development* **137,** 4135–4145 (2010).
31. Veenstra, J. A., Agricola, H. J. & Sellami, A. Regulatory peptides in fruit fly midgut. *Cell Tissue Res.* **334,** 499–516 (2008).
32. Housden, B. E. *et al.* Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Sci. Signal.* **8,** rs9 (2015).
33. Housden, B. E., Hu, Y. & Perrimon, N. Design and generation of *Drosophila* single guide RNA expression constructs. *Cold Spring Harb. Protoc*. http://doi.org/10.1101/pdb.prot090779 (2016).
34. Housden, B. E., Lin, S. & Perrimon, N. Cas9-based genome editing in *Drosophila*. *Methods Enzymol*. **546,** 415–439 (2014).
35. Ren, X. *et al.* Optimized gene editing technology for *Drosophila melanogaster* using germ line-specific Cas9. *Proc. Natl Acad. Sci. USA* **110,** 19012–19017 (2013).
36. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11,** 338–346 (2014).
37. Zhao, Y. *et al.* An expanded palette of genetically encoded Ca^{2+} indicators. *Science* **333,** 1888–1891 (2011).
38. Dai, W. & Montell, D. J. Live imaging of border cell migration in *Drosophila*. *Methods Mol. Biol*. **1407,** 153–168 (2016).
39. Wen, Q. *et al.* Proprioceptive coupling within motor neurons drives *C. elegans* forward locomotion. *Neuron* **76,** 750–761 (2012).
40. McDonald, J. C. *et al.* Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis* **21,** 27–40 (2000).
41. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
42. Micchelli, C. A., Sudmeier, L., Perrimon, N., Tang, S. & Beehler-Evans, R. Identification of adult midgut precursors in *Drosophila*. *GEP* **11,** 12–21 (2011).
43. Wang, C., Guo, X., Dou, K., Chen, H. & Xi, R. Ttk69 acts as a master repressor of enteroendocrine cell specification in *Drosophila* intestinal stem cell lineages. *Development* **142,** 3321–3331 (2015).



Extended Data Figure 1 | See next page for caption.

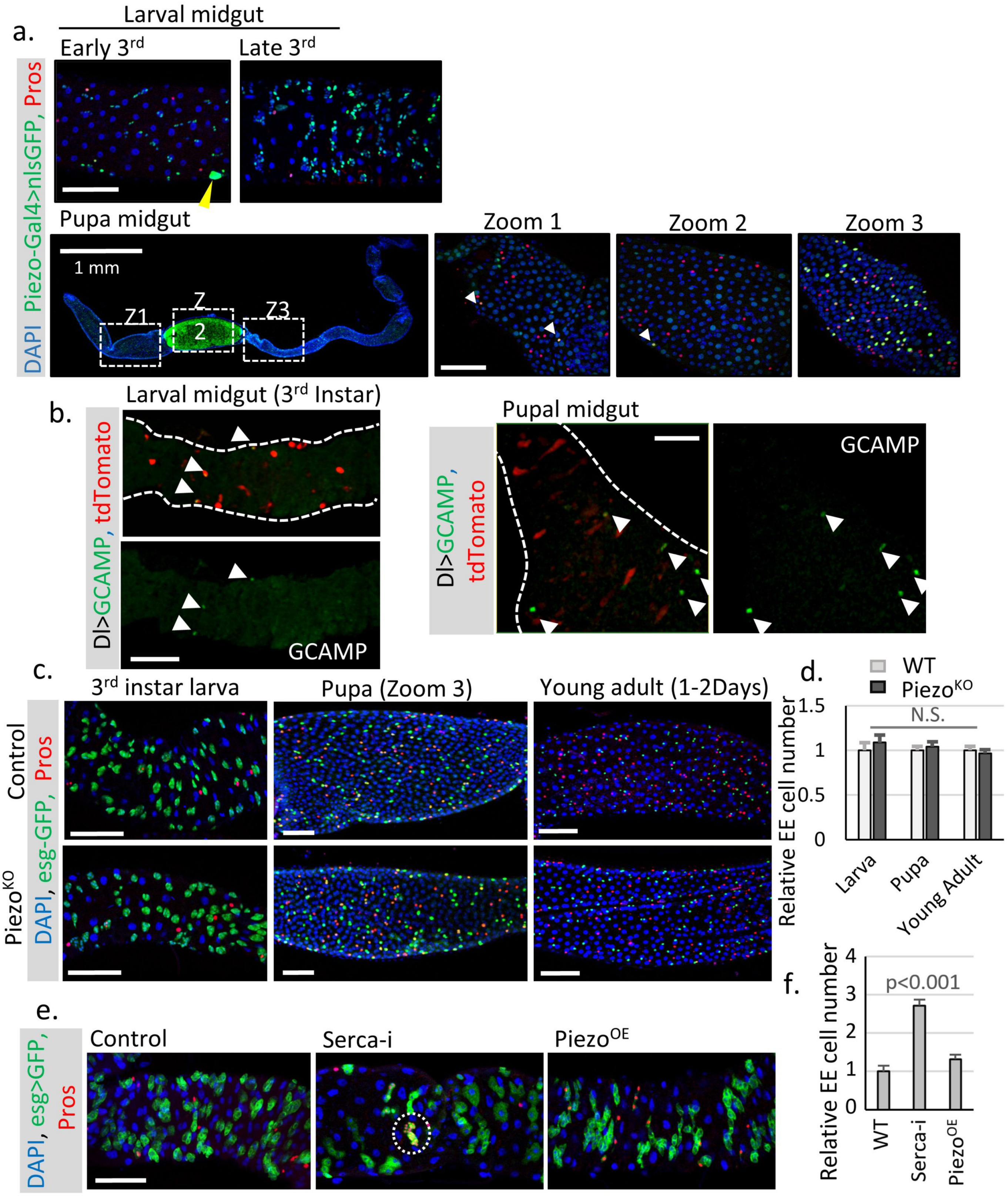
Extended Data Figure 1 | Piezo expression pattern and Piezo$^+$ cell lineage in the fly midgut. a, Expression pattern of GAL4 (BL59266) driven by the *Piezo* promoter[6]. **b**, Schematic of *Drosophila Piezo* gene structure. *GAL4* and the polyA tail were knocked in after the first start codon of *Piezo*; we refer to this knock-in GAL4 line as *Piezo-GAL4*. The ten predicted *Piezo* isoforms share the same N terminus. **c–f**, *Piezo* expression pattern in the midgut (*Piezo-GAL4*, *UAS-tdTomato3XHA*). Tissue was stained with an anti-haemagglutinin (HA) antibody to enhance the original signal. In addition to the small diploid stem cells, *Piezo* is also expressed in enterocytes after the cardia and around the copper and iron regions of the midgut. GAL4 activity outside the intestinal epithelium from tracheal cells can also be detected. **g**, Expression pattern of *Piezo* mRNA along different sections of the midgut. **h**, *Drosophila* midgut with Piezo$^+$ cells labelled by mCherry (*Piezo-GAL4*, *UAS-mCherry*; red) and Esg$^+$ cells labelled by *esg-GFP* (green). **i**, Midgut with Piezo$^+$ cells labelled by GFP (*Piezo-GAL4*, *UAS-Piezo-GFP*; green). Dl$^+$ stem cells were stained with an anti-Dl antibody (red). Arrowheads denote Piezo cells. **j**, Midgut expressing Piezo (GFP$^+$, green) in Esg$^+$ cells, with F-actin labelled by *UASp-Act5C-mRFP* (red). Piezo may form large cytoplasmic aggregates under stressed conditions[19], however, in the fly midgut, the GFP-tagged Piezo protein is localized primarily on the plasma membrane under both quiescent and over-proliferation conditions (**i**, **j**). **k**, *esg-GFP* is used as an indicator of newborn EEs. Under normal physiological conditions, around 2–3% of Esg$^+$ cells are also positive for Pros, suggesting that they are either differentiating or have just differentiated into EEs (denoted by arrowheads). All the newborn EEs are also positive for Piezo. Piezo and Pros double-positive but Esg-negative cells can be found occasionally (yellow arrowhead), most probably reflecting their late stage of differentiation. **l**, Piezo$^+$ newborn EEs are composed of both tachykinin-positive (Tk$^+$) and Tk$^-$ cells, suggesting that Piezo$^+$ cells are precursors for different types of EE. Arrowheads denote cells positive for both Piezo and Pros (left), Piezo and Tk (middle) or Pros and Tk (right). **m**, Dl$^+$, Su(H) Gbe$^+$ and Piezo$^+$ cells were traced using *Dl-GAL4*, *Su(H)Gbe-GAL4* and *Piezo-GAL4*. Arrowheads denote Pros (red) and GFP double-positive cells. **n**, Compared with *Dl-GAL4^{TS}*, which generates large GFP$^+$ enterocyte clones, *Piezo-GAL4^{TS}* primarily generates individual GFP$^+$ cells, with the occasional GFP$^+$ enterocyte cell clone (arrowhead). **o**, To visualize cells with GAL4 activity, which is repressed by the presence of *tub-GAL80^{TS}*, we incubated flies at 32 °C overnight before analysis. In this panel, two Pros$^+$ cells are GFP-positive but RFP-negative (indicated by arrowheads), suggesting that they are derived from Piezo$^+$ cells and then stop expressing Piezo. All experiments were independently repeated at least twice with similar results. Scale bars, 50 μm (**a**, **h**, **n**); 500 μm (**c**), 100 μm (**d–f**), 25 μm (**i**, **j**), 20 μm (**k–m**), 10 μm (**o**).



	Total	Piezo
🟦 Control	3.1	0.4
🟧 Bleo+	36.3	1.9

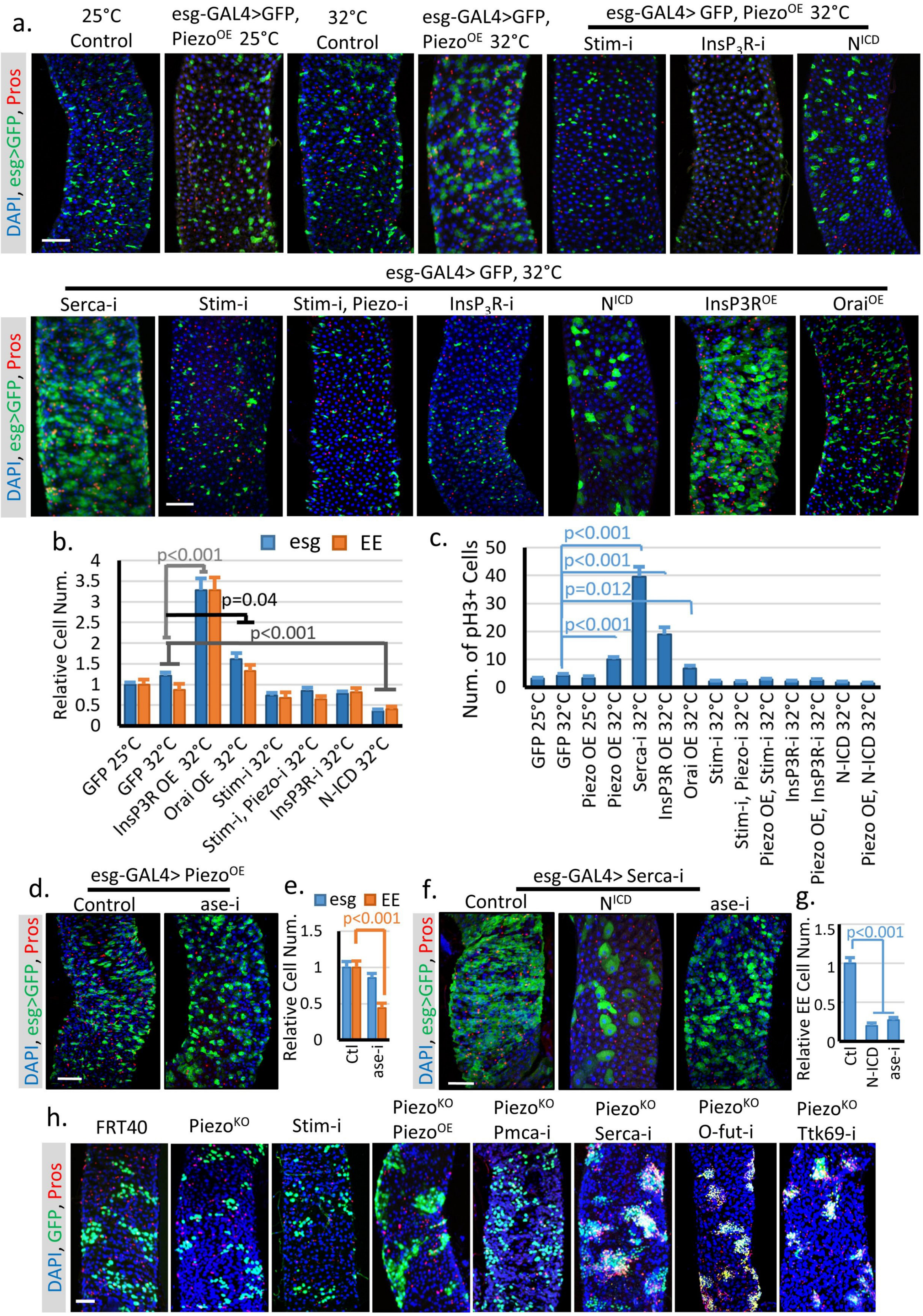
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Piezo$^+$ enteroendocrine precursors are ISC-derived EE precursors with reduced mitotic ability. a, Midguts from flies treated with bleomycin (10 μg ml^{-1} in 5% sucrose) or the γ-secretase inhibitor DAPT (4 mM in 5% sucrose). Arrowheads denote cells positive for both Piezo and Pros. Most (>95%) Piezo and Pros double-positive cells are also positive for Esg, suggesting that these cells are newborn EEs that still retain the *esg-GFP* signal. **b**, Percentage of newborn EEs (Piezo and Pros double-positive cells versus total Pros$^+$ EEs) in fly midguts under control, bleomycin and DAPT treatments. Cells within 200 μm × 200 μm areas, $n = 27$ (control), $n = 25$ (bleo), and $n = 22$ (DAPT), were analysed. **c**, Midgut with stem cells labelled by *esg-GFP* (green), Piezo$^+$ cells labelled by RFP (red), and mitotic cells labelled by anti-pH3 (magenta; arrowhead). **d**, **e**, Representative images of midguts from flies fed on either control (5% sucrose) or bleomycin (5% sucrose plus 10 μg ml^{-1} bleomycin) food. Piezo$^+$ enteroendocrine precursor cells are labelled by GFP (green), mitotic cells are labelled by pH3 staining (red). Arrowheads denote mitotic Piezo$^+$ cells. Because all pH3$^+$ cells are Dl$^+$ cells (according to the *Dl-lacZ*-labelled midgut), we counted all Piezo$^-$ pH3$^+$ cells as pH3$^+$ ISCs. Under both control (5% sucrose) and damage (5% sucrose + 10 μg ml^{-1} bleomycin) conditions, only around 8–10% of the pH3$^+$ cells are Piezo$^+$ (∼40% of total Dl$^+$ cells), suggesting that Piezo$^+$ cells are significantly less mitotically active than Piezo$^-$ Dl$^+$ cells. **f**, Around 50% of pH3$^+$ Piezo$^+$ cells show low levels of Pros staining. In addition, all pH3$^+$ Pros$^+$ cells are positive for Piezo, suggesting that Piezo$^+$ enteroendocrine precursor cells represent more general EE precursor cells than 'enteroendocrine mother cells' (EMCs)[21]. Arrowheads denote mitotic Piezo$^+$ cells. All experiments were independently repeated at least twice with similar results. **g**, Random GFP$^+$ clones were generated using *hsFLP*; *Ubi-*(*FRT.Stop*)*GFP*/*Piezo-GAL4*; *UAS-nlsRFP*. Flies (3–4 days old) were heat-shocked at 37 °C for 30 min once to induce clones in ISCs. The flies were then kept at 25 °C for 2 weeks before analysis. Within each GFP$^+$ clone, which is derived from ISCs, there are typically 1–2 Piezo$^+$ cells in the cluster (arrowheads), suggesting that Piezo$^+$ cells are generated from ISCs after adulthood. All experiments were independently repeated at least twice with similar results. Data are mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm (**a**, **c**), 20 μm (**d**, **f**) and 25 μm (**g**).



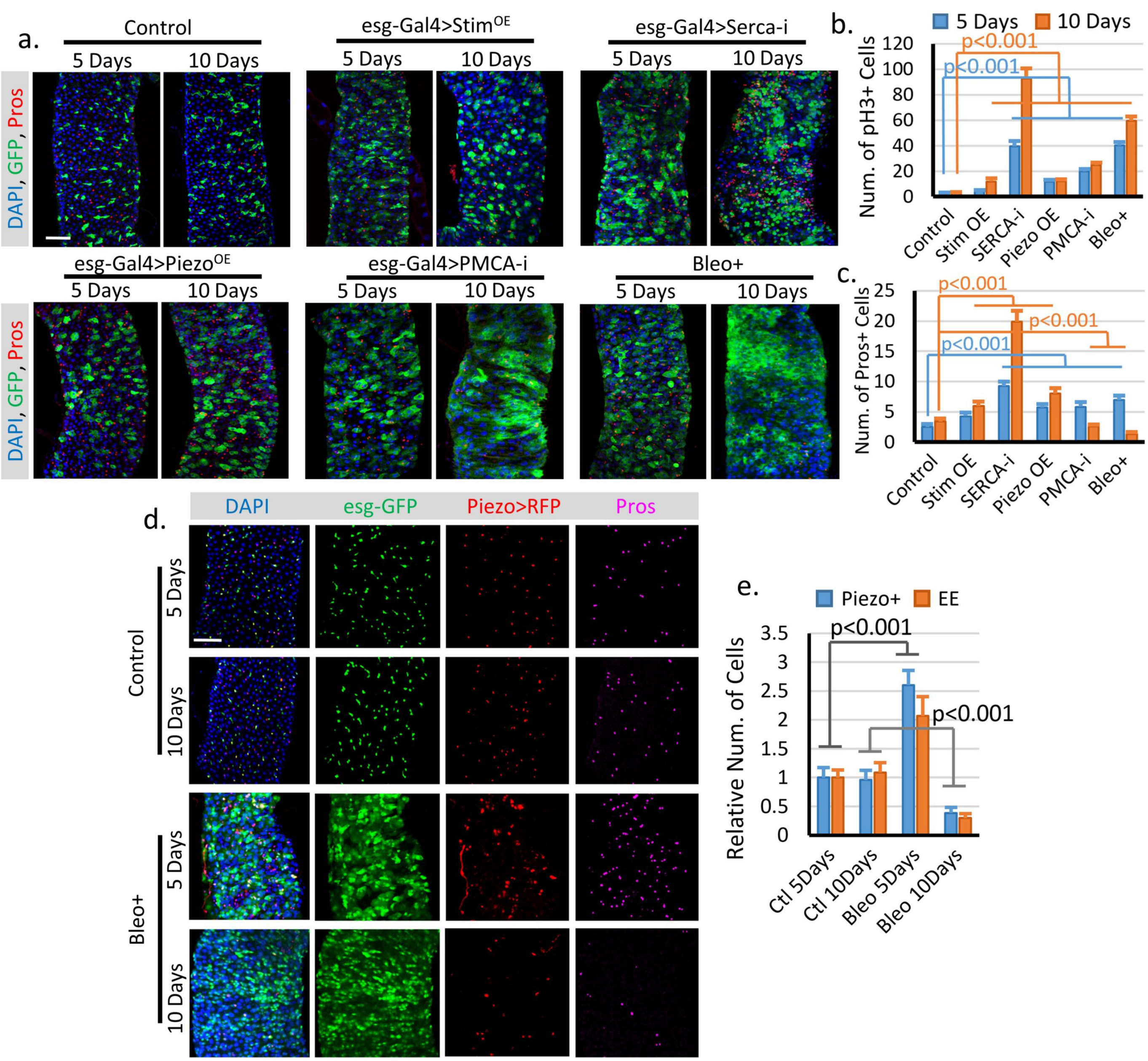
Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Expression and function of Piezo in larval and pupal midguts. a, Piezo$^+$ cells are labelled by GFP. Piezo is enriched in adult midgut precursor cells during larval stages. Strong expression of Piezo is also detected in tracheal cells associated with the midgut (yellow arrowhead denotes tracheal cell nucleus). After pupariation, the GFP signal can be detected at low levels in most midgut cells (including enterocytes), but is enriched in a few stem cells and EEs, which presumably are newborn EEs. Pupal gut 72 h after pupa formation is shown, with arrowheads denoting cells positive for both Piezo and Pros. High levels of Piezo are detected in a large number of EEs present in the pupal midgut, suggesting that the association of Piezo expression and EE differentiation is conserved during the pupal stage. **b**, Live imaging of larval and pupal midguts expressing GCAMP and tdTomato by *Dl-GAL4*. Arrowheads denote cells with high GCAMP activity. **c**, **d**, Midguts from *Piezo*-null (*Piezo^{KO}*) flies show no significant defects in EE generation during larval, pupal or early adult stages (1–2 days after eclosion). Number of midgut areas quantified: $n = 24$ (WT, larva), $n = 23$ (WT, pupa), $n = 28$ (WT, young adult), $n = 23$ (*Piezo^{KO}*, larva), $n = 23$ (*Piezo^{KO}*, pupa), $n = 28$ (*Piezo^{KO}*, young adult). These results indicate that mechanically controlled Piezo activation is not the major mechanism for EE production during early development. Unlike the adult midgut, the larval midgut does not regenerate through mitosis and only grows through increases in cell size. It is only during late stages of third instar larval development that the quiescent adult midgut precursor cells start to proliferate and generate both new enterocytes and EEs for pupal gut formation, and most new EEs (~several hundred) are created within a very narrow time window approximately 72–96 h after pupa formation[42]. Therefore, the generation of EEs is 15–30 times faster at that stage than during the adult stage under physiological condition, suggesting that a different mechanism that stimulates strong acute EE differentiation is involved during developmental stages. **e**, **f**, Knockdown of SERCA using *esg-GAL4* during larval stages significantly increases EE cell number. Conversely, overexpression of *Piezo* (*Piezo^{OE}*) has no significant phenotype. White circle denotes a cluster of extra EE cells. Number of midgut areas quantified: $n = 26$ (WT), $n = 28$ (*SERCA-i*), $n = 26$ (*Piezo^{OE}*). All experiments were independently repeated at least twice with similar results. Data are mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm.

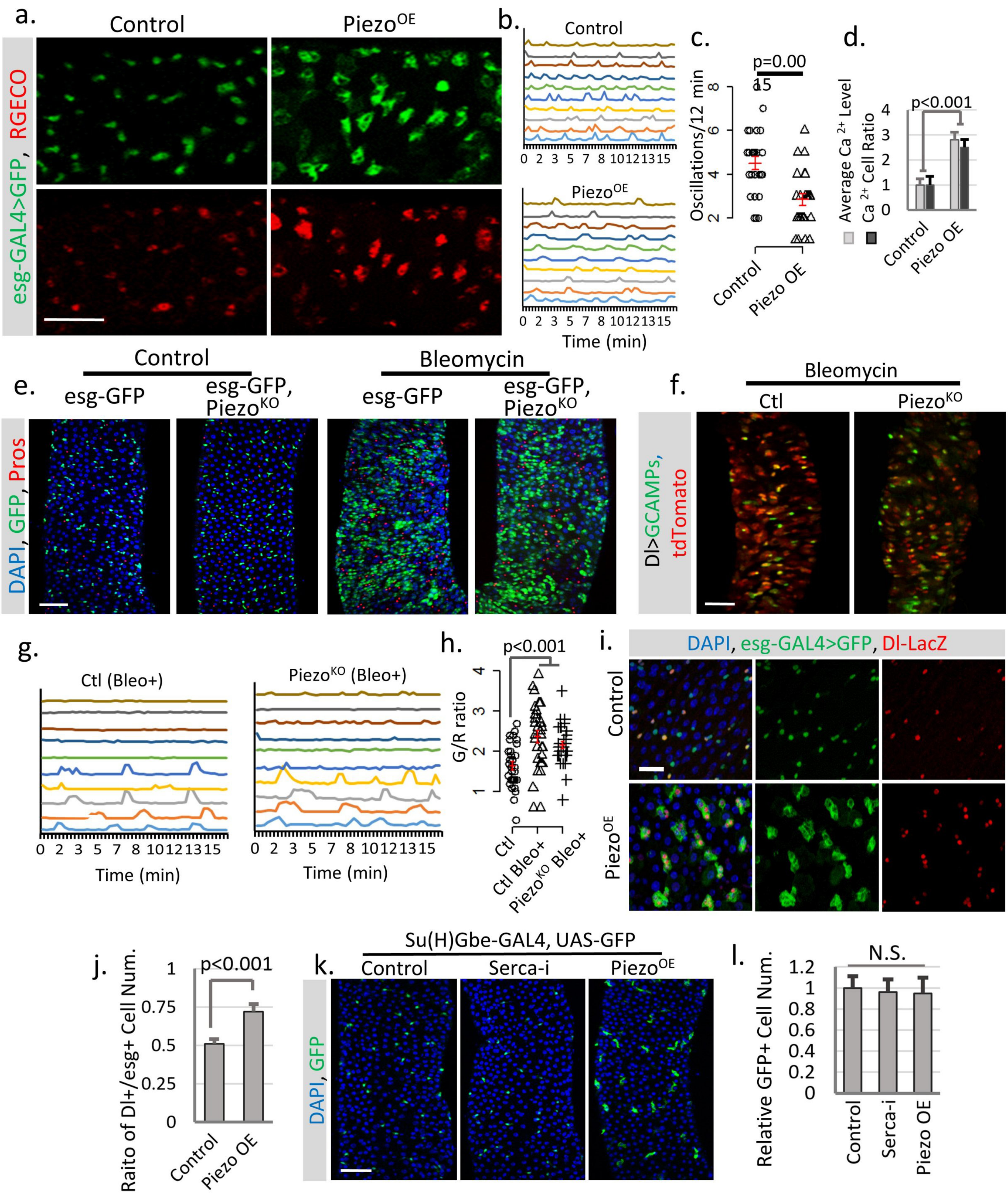


Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | Piezo regulates stem-cell differentiation primarily through Ca^{2+} signalling, which is upstream of Notch, Ttk69 and the *achaete-scute* gene complex (AS-C). a, Phenotypes associated with *UAS-GFP* (at 25 °C or 32 °C), *UAS-Piezo^{OE}* together with *Stim-i*, *InsP3R-i* and N^{ICD}, and *UAS-GFP* together with *Stim-i*, *Stim-i* + *Piezo-i*, *InsP3R-i*, N^{ICD}, *InsP3R* overexpression (*InsP3R^{OE}*), and *Orai^{OE}* (at 32 °C). Overexpression of *Piezo* using *esg-GAL4* did not show a significant phenotype at 25 °C. By contrast, incubation at 32 °C for 4 days showed an increased in the number of both Esg^+ cells and $Pros^+$ EEs. Moderate overexpression of *Piezo* at 25 °C had no significant effects. However, strong overexpression at 32 °C caused an increase in both Esg^+ cells and EEs, which phenocopied the increase of cytosolic Ca^{2+} through SERCA reduction. All flies were incubated at the indicated temperature for 4–5 days before analysis. **b**, The number of Esg^+ and $Pros^+$ cells within 10,000-μm^2 area. Number of midgut areas quantified: $n = 30$ (GFP 25 °C), $n = 31$ (GFP 32 °C), $n = 25$ (*InsP3R^{OE}* 32 °C), $n = 27$ (*Orai^{OE}* 32 °C), $n = 31$ (*Stim-i* 32 °C), $n = 27$ (*Stim-i*, *Piezo-i* 32 °C), $n = 29$ (*InsP3R-i* 32 °C), $n = 29$ (N^{ICD} 32 °C). **c**, Average number of mitotic cells within the fly midgut from indicated genotypes. Number of midguts analysed: $n = 20$ (GFP 25 °C), $n = 19$ (GFP 32 °C), $n = 20$ (*Piezo^{OE}* 25 °C), $n = 19$ (*Piezo^{OE}* 32 °C), $n = 18$ (*SERCA-i*, 32 °C), $n = 18$ (*InsP3R^{OE}* 32 °C), $n = 24$ (*Orai^{OE}* 32 °C), $n = 19$ (*Stim-i,* 32 °C), $n = 19$ (*Stim-i*, *Piezo-i* 32 °C), $n = 19$ (*Piezo^{OE}*, *Stim-i*, 32 °C), $n = 18$ (*InsP3R-i* 32 °C), $n = 18$ (*Piezo^{OE}*, *InsP3R-i* 32 °C), $n = 17$ (N^{ICD} 32 °C), $n = 17$ (*Piez^{OE}*, N^{ICD} 32 °C). **d**, **e**, EE production induced by overexpression of *Piezo* is blocked by RNAi that targets the *acheate-scute* complex (AS-C) component *asense* (*ase*). Number of midgut areas quantified: $n = 29$ (ctl), $n = 30$ (*ase-i*). **f**, **g**, Expression of N^{ICD} in the presence of *SERCA-i* significantly reduced both stem-cell proliferation and EE production. Knockdown of *ase* specifically blocks EE differentiation but not proliferation. Number of midgut areas quantified: $n = 27$ (ctrl), $n = 24$ (N^{ICD}), $n = 25$ (*ase-i*). Even though *ttk69* (also known as *ttk*) and *AS-C* knockdown affect Piezo- and SERCA-related phenotypes, Ca^{2+} signalling probably does not directly affect Ttk69 or *AS-C*; previous studies have shown that Ttk69 and AS-C reduction can convert Notch-high enteroblasts into EEs[43], but neither *Piezo* overexpression nor *SERCA* knockdown has any effect in enteroblasts. **h**, MARCM clones of cells homozygous for FRT (control), *Piezo^{KO}*, *Stim-i*, *Piezo^{KO}* + *Piezo^{OE}*, *Piezo^{KO}* + *PMCA-i*, *Piezo^{KO}* + *SERCA-i*, *Piezo^{KO}* + *O-fut1-i*, and *Piezo^{KO}* + *ttk69-i*. Rescue/reversion of the reduction of EEs in *Piezo*-null clones by increasing levels of cytosolic Ca^{2+} (by knocking down the Ca^{2+} export pump PMCA or endoplasmic reticulum Ca^{2+} ATPase SERCA) or by reducing Notch activity (by knocking down its key processing enzyme O-fut1, and knocking down EE cell fate repressor Ttk69). All data are from at least two independent replicates and are expressed as mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 µm.

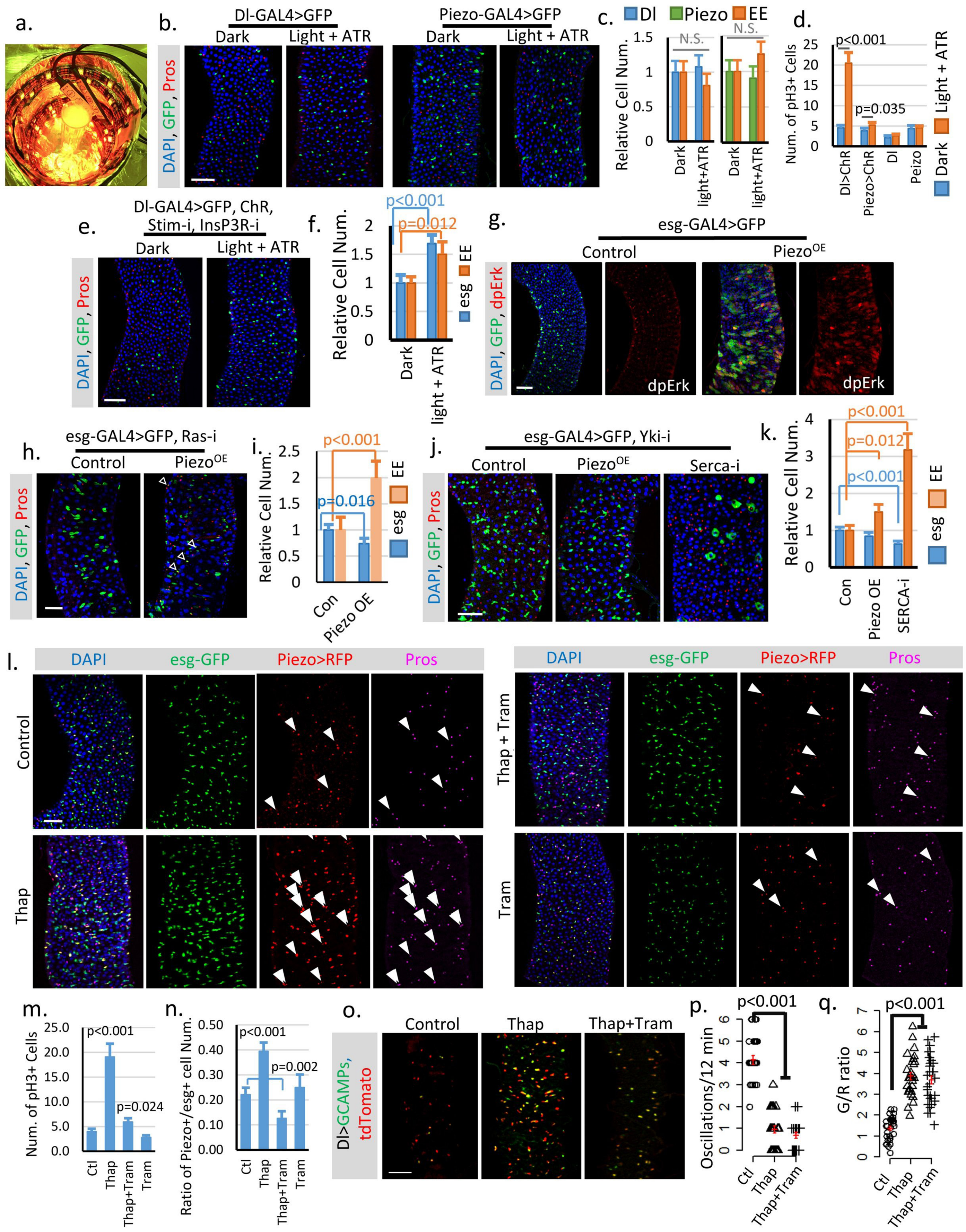


Extended Data Figure 5 | Prolonged increase of stem-cell proliferation may reduce EE cell number. a, Fly midguts of each indicated genotype/condition were analysed after incubation for 5 and 10 days at 32 °C. Esg^{+} cells (GFP^{+}, green) and EE cells (Pros^{+}, red). Representative images from two independent replicates. **b**, Quantification of mitosis (pH3^{+} cell number) of midguts from flies expressing GFP only (control, $n = 16$ (5 days), $n = 16$ (10 days)), full-length *Stim* (*Stim^{OE}*, $n = 15$ (5 days), $n = 17$ (10 days), *SERCA-i* ($n = 18$ (5 days), $n = 16$ (10 days)), *Piezo^{OE}* ($n = 17$ (5 days), $n = 18$ (10 days)), *PMCA-i* ($n = 15$ (5 days), $n = 15$ (10 days)), and flies fed bleomycin-containing food (regular food + 10 μg ml^{-1} bleomycin, $n = 15$ (5 days), $n = 13$ (10 days)). **c**, Quantification of Pros^{+} EE cell number from 10,000-μm^{2} regions: $n = 31$ (5 days), $n = 30$ (10 days) (control); $n = 30$ (5 days), $n = 32$ (10 days) (*Stim^{OE}*); $n = 30$ (5 days), $n = 30$ (10 days) (*SERCA-i*); $n = 31$ (5 days), $n = 32$ (10 days) (*Piezo^{OE}*); $n = 32$ (5 days), $n = 31$ (10 days) (*PMCA-i*); $n = 29$ (5 days), $n = 28$ (10 days) (Bleo+). Bleomycin treatment or *PMCA-i* significantly reduced the number of EEs. This reduction is primarily due to increased turnover of EEs, as blocking cell mitosis for 5 days had no significant effect on EE cell number (Extended Data Fig. 7). The differences between stem-cell proliferation and EE differentiation may be due to a different level of cytosolic Ca^{2+} increase and the Ca^{2+} depletion in the ER store. **d**, Change of Piezo^{+} cells and EEs after 5 and 10 days of control (5% sucrose) or bleomycin (5% sucrose plus 10 μg ml^{-1} bleomycin) treatment. Representative images from two independent replicates. **e**, Quantification of Piezo^{+} cells and EEs from 10–15 midguts for each condition. Both Piezo^{+} cells and EEs number increased after 5 days of bleomycin treatment, and significantly decreased after 10 days of treatment. Cell numbers were quantified within a 10,000-μm^{2} area, except for pH3, which is quantified from the whole midgut. All data are mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm.



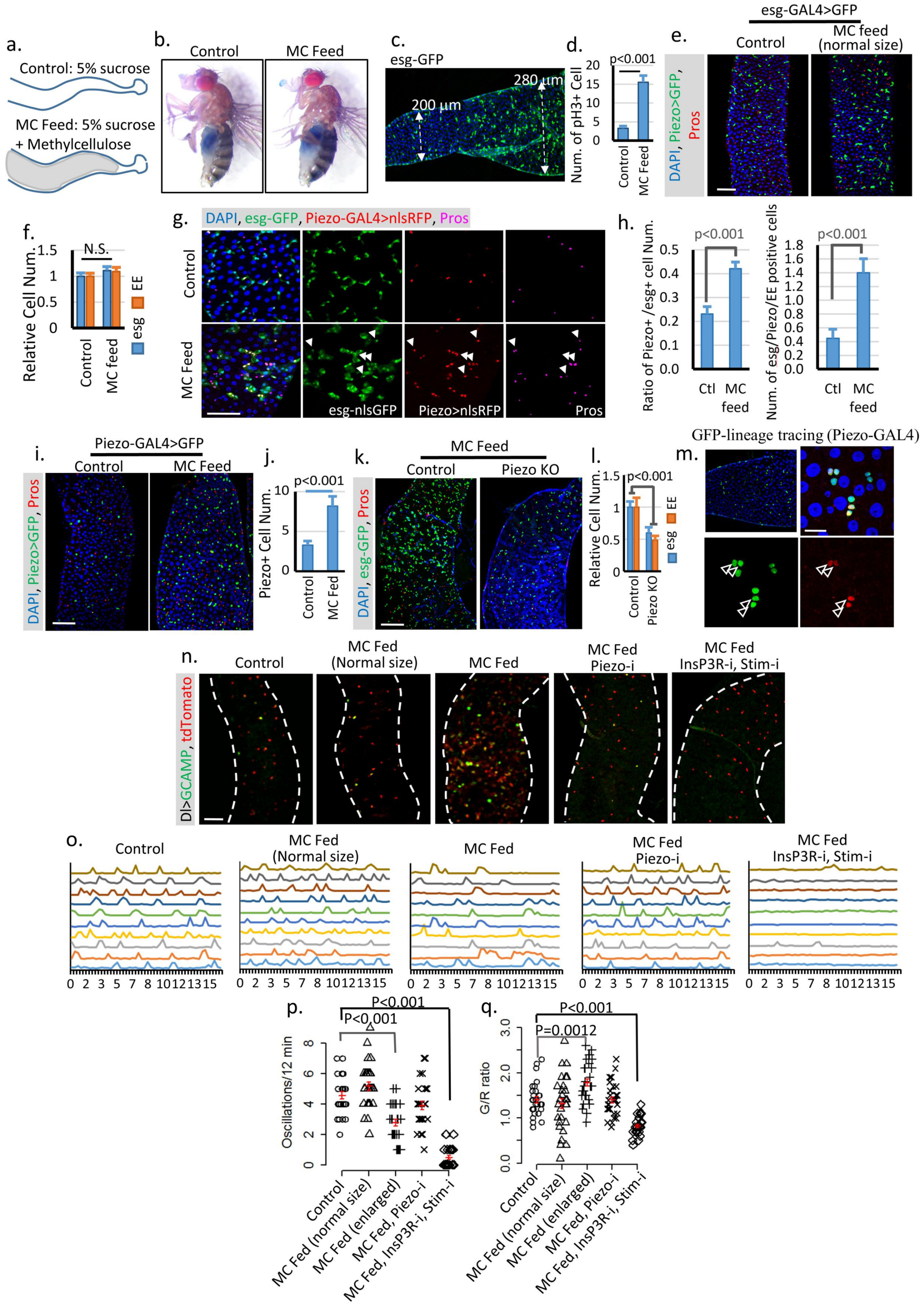
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | *Piezo* overexpression increases cytosolic Ca^{2+} levels, which further triggers proliferation of ISCs but not enteroblasts. **a**, Overexpression of *Piezo* in Esg^+ (GFP^+, green) cells at 32 °C causes an increase in cytosolic Ca^{2+} (indicated by the red fluorescent calcium reporter RGECO) compared to control (*esg-GAL4/UAS-GFP*, *UAS-RGECO*). Representative images from three short time-lapse imagings of cultured fly midguts. Scale bar, 50 μm. **b**, Typical traces of Ca^{2+} oscillations in Esg^+ cells of midgut from either control or $Piezo^{OE}$ flies from three independent replicates. **c**, Ca^{2+} oscillation frequency of Esg^+ cells from either control or *Piezo* overexpression midguts. Data are from 27 cells from three replicates for each condition. **d**, Average RGECO signal intensity in all GFP^+ cells (blue) and percentage of Ca^{2+}-positive cells (signal higher than 3× the s.d. of background) compared to total GFP^+ cells (orange). Signal intensities were calculated from 10,000-μm^2 regions: $n = 17$ (control), $n = 22$ ($Piezo^{OE}$) from three independent experiments. **e**, Bleomycin (10 $\mu g\ ml^{-1}$) (5 days of treatment) triggers a significant increase in Esg^+ cells and EE cells in both wild-type and $Piezo^{KO}$ flies. Representative images from three independent replicates are shown. **f**, Images of live midguts from wild-type and $Piezo^{KO}$ flies. Flies were fed on food containing bleomycin for 3 days before imaging. **g**, **h**, Traces of Ca^{2+} oscillations in Dl^+ stem cells from wild-type and *Piezo* mutant flies fed on bleomycin for 4–5 days. Bleomycin treatment causes some stem cells to maintain constant high Ca^{2+} levels, whereas others show reduced oscillation frequency but an increased average GCaMP/RFP intensity (G/R) ratio. These data show that tissue damage by bleomycin triggers stem-cell proliferation, EE production and an increase in cytosolic Ca^{2+}, independently of Piezo. Thirty cells from $n = 4$ (control), $n = 4$ (Bleo+), and $n = 5$ ($Piezo^{KO}$ and Bleo+) independent guts are plotted. **i**, Overexpression of $Piezo^{OE}$ in Esg^+ cells (32 °C) increases the proportion of Dl^+ cells (labelled by *Dl-lacZ*; red) within the Esg^+ population. **j**, *Piezo* overexpression promotes the Dl^+/Esg^+ cell ratio. Ratio between Dl^+ and Esg^+ cells within 10,000-μm^2 regions: $n = 21$ (control) and $n = 22$ ($Piezo^{OE}$) from two independent replicates, are analysed. **k**, **l**, Overexpression of *Piezo* or knockdown of *SERCA* in $Su(H)Gbe^+$ enteroblast cells showed no significant phenotype, suggesting that their effect may be blocked by high Notch activity. Number of midgut areas quantified: $n = 18$ (control), $n = 20$ (*SERCA-i*), $n = 16$ ($Piezo^{OE}$). Data are mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm (**a**, **e**, **f**, **k**) and 20 μm (**i**).



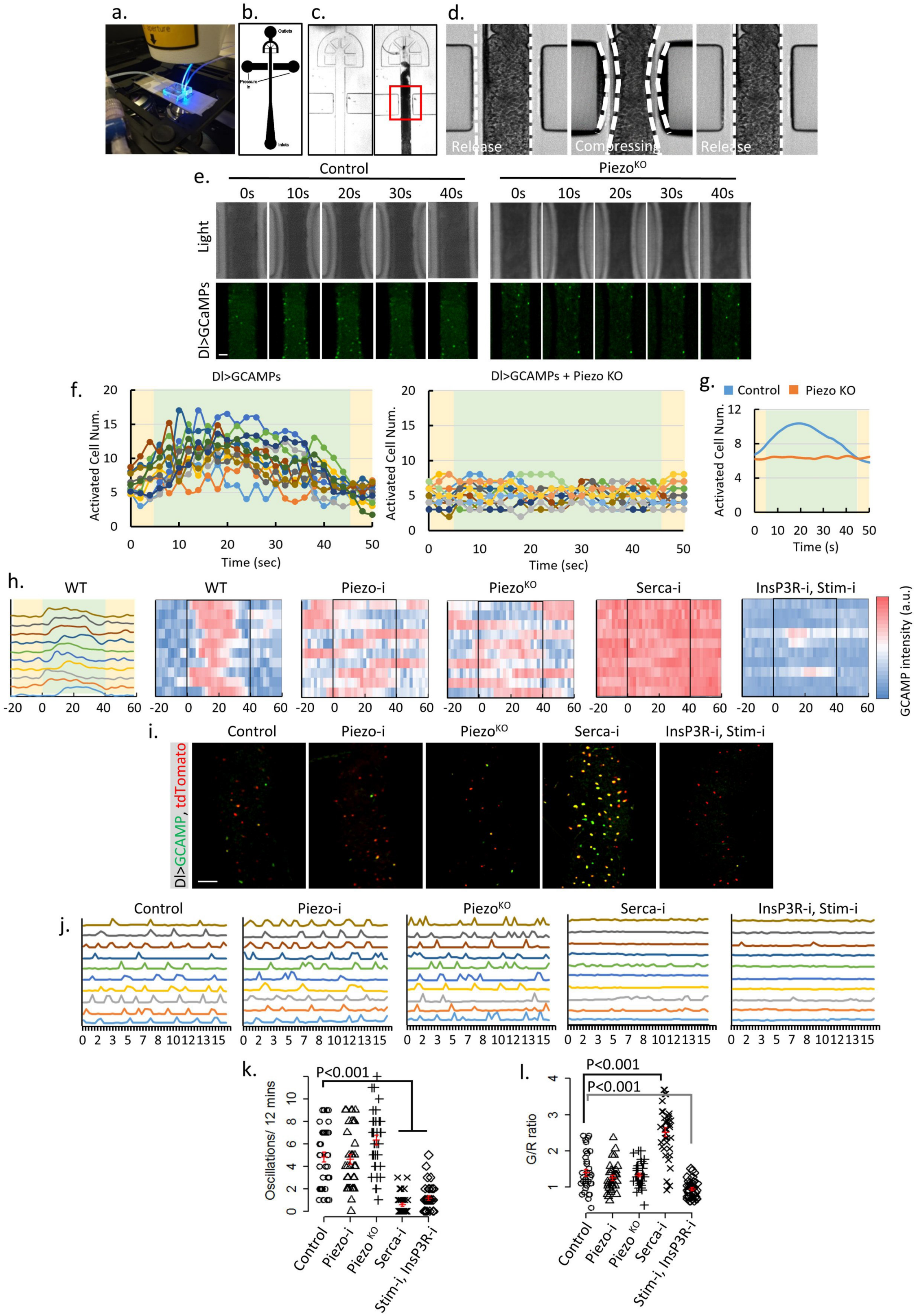
Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Cytosolic Ca^{2+} triggers ISC proliferation and enteroendocrine precursor differentiation into EEs. a, Image of chamber used for optogenetic activation of ChR. **b**, **c**, Flies expressing GFP only in Dl^{+} stem cells or Piezo^{+} enteroendocrine precursor (EE precursor) cells were treated under either dark or light + ATR conditions for 2 weeks, as per the flies expressing ChR. No significant phenotype was induced by the treatment alone. Number of midgut areas quantified: $n = 29$ (Dl, dark), $n = 33$ (Dl, light + ATR), $n = 31$ (Piezo, dark), $n = 34$ (Piezo, light + ATR). Representative results from two independent replicates are shown. **d**, Mitosis quantification of midgut from indicated genotype/condition. Activating ChR in Dl^{+} cells significantly promotes stem-cell proliferation. Only a mild increase in mitosis was detected in ChR-active Piezo^{+} enteroendocrine precursor cells, suggesting that the primary effect of Ca^{2+} in enteroendocrine precursor cells is to promote differentiation. Data are from 30 guts (*Dl>ChR*); 30 guts (*Piezo>ChR*); 29 guts (*Dl*); guts (*Piezo*) from two independent replicates. pH3^{+} cell number is quantified from the whole midgut. **e**, **f**, Activation of the channelrhodopsin CsChrimson in Dl^{+} stem cells with both *Stim* and *InsP3R* knocked down causes a reduced increase in stem cells and EEs compared to wild-type stem cells. Flies were raised at 18 °C and shifted to 25 °C during the experiment. Cell numbers are quantified within a 10,000-μm^{2} area from 29 regions (dark) and 31 regions (light + ATR) from two independent replicates. **g**, Overexpression of *Piezo* in Esg^{+} cells increases MAPK pathway activity. Phosphorylation of dpErk is significantly increased in *Piezo*-overexpressing cells. Representative images from two independent experiments are shown. **h**, **i**, Knockdown of Ras significantly reduces stem-cell proliferation caused by *Piezo* overexpression, but does not block Piezo-triggered EE differentiation. Flies were kept at 32 °C for 4–5 days before analysis. Esg^{+} and EE cell numbers were quantified from $n = 29$ (control) and $n = 30$ (*Piezo^{OE}*) midgut areas from two independent experiments. Arrowheads denote newborn EEs (positive for both Esg and Pros). **j**, **k**, Knockdown of *yorkie* using *yki-i* completely blocks stem-cell proliferation but not the increase of EE cells induced by either *Piezo* overexpression or *SERCA* knockdown. In addition, knockdown of *SERCA* together with *yki* also significantly reduced stem-cell number, suggesting a depletion of stem cells caused by constant EE differentiation. Cell numbers were quantified from 30 midgut areas for each genotype. **l**, Midguts from flies fed on control (5% sucrose), thapsigargin (5% sucrose, 0.5 μM thapsigargin; Thap), thapsigargin + trametinib (5% sucrose, 0.5 μM thapsigargin, 10 μM trametinib; Thap + Tram), and trametinib (5% sucrose, 5 μM trametinib; Tram) for 4 days. Representative images from three independent experiments are shown. The increase of cytosolic Ca^{2+} by thapsigargin promotes stem-cell proliferation, enteroendocrine precursor (Piezo^{+} cell) production, and EE differentiation. White arrowheads denote newborn EEs (positive for Esg, Piezo and Pros). **m**, Quantification of mitotic cells from $n = 15$ (control), $n = 16$ (Thap), $n = 17$ (Thap + Tram), and $n = 16$ (Tram) midguts. Thapsigargin treatment triggers a significant increase in mitosis, which is largely reduced by the MAPK inhibitor trametinib. **n**, Percentage of Piezo^{+} cells within the Esg^{+} cell population. Number of areas quantified: $n = 29$ (ctl), $n = 31$ (Thap), $n = 32$ (Thap + Tram), $n = 29$ (Tram). **o**, Representative Ca^{2+} images of live midgut from control, thapsigargin-treated, and thapsigargin plus trametinib-treated flies. Similar results were collected from 4 independent guts for each condition. **p**, **q**, Thapsigargin treatment caused a reduction in oscillation frequency but an increase in the average GCaMP/RFP (G/R) ratio. The increase in cytosolic Ca^{2+} by thapsigargin treatment is not affected by MAPK inhibition. Data are from 29 cells from 3 independent guts for each condition. Data are mean + s.e.m. (shown in red). *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm.



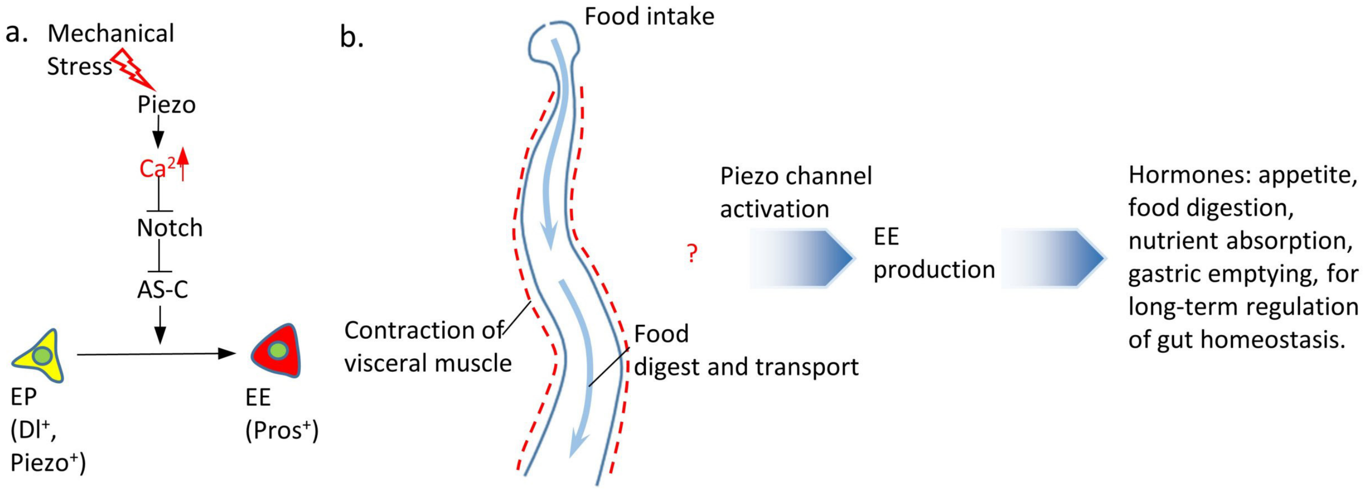
Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Over-feeding triggers stem-cell proliferation and an increase in EEs. a, Schematic illustration of fly midguts from control (5% sucrose) or methylcellulose (5% sucrose plus 10% methylcellulose) fed flies. **b**, 'Smurf' assay of flies fed on both control and methylcellulose food shows no damage to gut integrity. Two independent replicates showed similar results. **c**, **d**, Image of a midgut of a fly fed on methylcellulose food. The cell proliferation phenotype is associated with an increase in midgut diameter but not food content. Data are from 23 midgut areas from 2 independent experiments for each condition. **e**, **f**, Midguts from flies fed methylcellulose with no increase in gut diameter show no change in phenotype compared with control. Data are from 31 regions (control) and 28 regions (MC) from three independent experiments. **g**, **h**, Feeding-induced cell proliferation produces more Piezo$^+$ cells, which differentiate into EEs. White arrowheads denote newborn EEs. Data are from 27 areas from 2 independent experiments for each condition. **i**, **j**, Feeding-induced midgut enlargement triggers a significant increase in the enteroendocrine precursor/Piezo$^+$ cell number. Data are from $n = 30$ (control) and $n = 32$ (MC) midgut areas from 2 independent replicates. **k**, **l**, Feeding-trigged stem-cell proliferation and EE increases are blocked in the *Piezo*-null (*Piezo*^{KO}) mutant. Data are from $n = 27$ (control) and $n = 32$ (*Piezo*^{KO}) midgut areas from 2 independent replicates. **m**, Linage-tracing experiment (using *Piezo-GAL4*) under overfed conditions shows a significant increase in cell number (2–3) in the same cluster compared to tracing result under control conditions, suggesting that either more Piezo cells were created from ISCs or more Piezo$^+$ cells divide to create more progeny. Arrowheads denote cells positive for both GFP and Pros. **n**, Images of live midguts from the following conditions/genotypes: control, methylcellulose fed without midgut diameter increase (normal size), methylcellulose fed with enlarged midgut diameter, methylcellulose fed with *Piezo-i* and enlarged midgut diameter, and methylcellulose fed with *InsP3R-i* + *Stim-i* and enlarged midgut diameter. **o**, Representative traces of Ca^{2+} oscillations in Dl$^+$ stem cells of flies from indicated treatment/genotypes. Data are from 3 independent experiments for each genotype/condition. **p**, **q**, Ca^{2+} oscillation frequency and GCaMP/RFP intensity ratio of 30 cells from three individual guts for each genotype. Mean ± s.e.m. is displayed in red. Enlarged midgut of fly fed on methylcellulose shows reduced Ca^{2+} oscillation frequency but increased average cytosolic Ca^{2+} level. Methylcellulose alone does not trigger any significant change in Ca^{2+} activity. Knockdown of *Piezo* or of both *Stim* and *InsP3R* blocks this feeding-induced increase in cytosolic Ca^{2+}. Knockdown of *InsP3R* or *Stim* alone has no significant effect on cytosolic Ca^{2+} (data not shown), which is probably due to the reduced expression levels of *Dl-GAL4* compared with *esg-GAL4*. The change in Ca^{2+} activity in enlarged midguts of methylcellulose-fed flies is similar to some cells in the bleomycin-damaged midguts (Extended Data Fig. 6f, g). However, most cells from enlarged midguts of methylcellulose-fed flies still oscillate, which is different from stem cells in bleomycin-treated midguts, in which a large portion of cells maintain a constant high level of Ca^{2+} (Extended Data Fig. 6f, g). Data are mean + s.e.m. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm (**e**, **i**, **k**, **n**), 25 μm (**g**) and 10 μm (**m**).



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Direct mechanical activation of the Piezo channel triggers an increase in cytosolic Ca^{2+} in stem cells. a, Image of the microfluidic chip used for the *ex vivo* mechanical trigger experiment. **b**, **c**, Design of the channels on the microfluidic chip. Compressed air was delivered through the left and right channels and controlled by a manual gauge. Dissected fly midguts were loaded into the main channel (centre) from an inlet at the bottom. **d**, During each compression cycle, the midgut was squeezed to achieve an approximately 30–35% reduction in diameter from both sides. The switching time between compression and relaxation is approximately 1 s. **e**, Representative samples of *ex vivo* mechanical trigger experiment. Time 0 s and 40 s were taken immediately before and after compression. The total compression time is 40 s. Transmission light (top) and GCaMP6s signal (bottom) are shown. Compared to control, the loss of Piezo significantly blocked activation of stem cells by mechanical compression. **f**, Plots of activated cells numbers during one triggering cycle (50 s) for control ($n = 12$) and $Piezo^{KO}$ ($n = 15$) fly midguts. Data were from 4–5 individual midguts. All GCaMP-positive cells (5-fold brighter than the midgut autofluorescence signal) within the field were counted. Periods of compression and relaxation are indicated by green and yellow colours, respectively. **g**, Averaged response curves of multiple compression cycles ($n = 12$ for control and $n = 10$ for $Piezo^{KO}$) from control (blue) and $Piezo^{KO}$ (orange) midguts. **h**, Typical traces of Ca^{2+} activities in wild-type stem cells that respond to the mechanical stimulus. Data are represented in curve plot (first panel) and heat-map plot (second panel). The compression period is from 0 to 40 s (black box). Typical traces of Ca^{2+} activities with indicated genotypes. Stem cells with *Piezo* knockdown or mutant do not respond to the mechanical stimulus. Knockdown of *SERCA* causes a constant high cytosolic Ca^{2+}. Knockdown of both *Stim* and *InsP3R* significantly reduces random Ca^{2+} activities and largely blocks the mechanically triggered Ca^{2+} increase. Data are from three independent experiments for each genotype/condition. **i**, Images of cultured midguts from control, *Piezo-i*, $Piezo^{KO}$, *SERCA-i*, and *InsP3R-i* + *Stim-i* flies. **j**, Typical traces of Ca^{2+} activities in stem cells of indicated genotypes. Data are from three independent guts for each genotype/condition. **k**, **l**, Ca^{2+} oscillation frequency and GCaMP/RFP (G/R) intensity ratio in $n = 35$ cells (control), $n = 35$ cells (*Piezo-i*), $n = 34$ cells ($Piezo^{KO}$), $n = 36$ cells (*SERCA-i*), $n = 33$ cells (*InsP3R-i* + *Stim-i*) from three independent experiment for each condition/genotype. Neither *Piezo-i* nor $Piezo^{KO}$ significantly affect Ca^{2+} activities. Knockdown of *SERCA* induces a constant increase of cytosolic Ca^{2+} in most cells. Knockdown of both *InsP3R* and *Stim* stem cells significantly reduces their Ca^{2+} activities. Our data indicate that mechanical stresses generated during food digestion may activate Piezo and promote EE generation *in vivo*. However, the timescale between our *ex vivo* mechanical activation and *in vivo* cell proliferation and differentiation experiment is very different, especially as the *in vivo* property of Piezo-mediated Ca^{2+} activity in enteroendocrine precursor cells is unknown. According to our observations, only a small percentage (<5%) of $Piezo^{+}$ cells become EEs every day under normal conditions (interpreted from the Piezo and Pros double-positive cell number). Therefore, it is possible that either Piezo is difficult to activate *in vivo* by physiological levels of mechanical stimuli, or long-term cumulative Piezo activation is required to trigger EE differentiation. Mean ± s.e.m. is displayed in red. *P* values are from a two-tailed Student's *t*-test with unequal variance. Scale bars, 50 μm.


a.
Mechanical Stress
Piezo
Ca2+
Notch
AS-C
EP
(Dl+, Piezo+)
EE
(Pros+)
b.
Food intake
Contraction of visceral muscle
Food digest and transport
?
Piezo channel activation
EE production
Hormones: appetite, food digestion, nutrient absorption, gastric emptying, for long-term regulation of gut homeostasis.

Extended Data Figure 10 | A model of Piezo activation and downstream signalling. a, Under normal conditions, Piezo^{+} cells, which we refer to as endocrine precursor (EP) cells, are unipotent stem cells that are mitotically quiescent and have a predetermined EE cell fate. In the presence of mechanical stimulation, the Piezo channel is activated and leads to an increase in cytosolic Ca^{2+} in Piezo^{+} enteroendocrine precursor cells. Ca^{2+} increases in enteroendocrine precursor cells trigger strong cell differentiation into EEs, which is probably mediated by inhibition of Notch activity and consequent increase of AS-C transcription activity. **b**, The presence of food in the intestine triggers increased mechanical stress during food transport and visceral muscle contraction. Our results suggest that mechanical signalling activates the mechanosensitive channel Piezo in quiescent enteroendocrine precursor cells and leads to an increase in cytosolic Ca^{2+} levels, which maintain the basal level EE cell production under physiological conditions and promote fast EE generation under abnormally fed conditions. We hypothesize that, as a key regulator of midgut function, EE cells might secrete hormones to enhance different long-term gastric functions including appetite, digestion, nutrient absorption or gastric emptying.

LETTER

doi:10.1038/nature25757

Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells

Yoav Lubelsky[1] & Igor Ulitsky[1]

Long noncoding RNAs (lncRNAs) are emerging as key parts of multiple cellular pathways[1], but their modes of action and how these are dictated by sequence remain unclear. lncRNAs tend to be enriched in the nuclear fraction, whereas most mRNAs are overtly cytoplasmic[2], although several studies have found that hundreds of mRNAs in various cell types are retained in the nucleus[3,4]. It is thus conceivable that some mechanisms that promote nuclear enrichment are shared between lncRNAs and mRNAs. Here, to identify elements in lncRNAs and mRNAs that can force nuclear localization, we screened libraries of short fragments tiled across nuclear RNAs, which were cloned into the untranslated regions of an efficiently exported mRNA. The screen identified a short sequence derived from Alu elements and bound by HNRNPK that increased nuclear accumulation. Binding of HNRNPK to C-rich motifs outside Alu elements is also associated with nuclear enrichment in both lncRNAs and mRNAs, and this mechanism is conserved across species. Our results thus identify a pathway for regulation of RNA accumulation and subcellular localization that has been co-opted to regulate the fate of transcripts with integrated Alu elements.

With a few exceptions, the detailed mechanisms and sequence elements that can direct nuclear enrichment of long RNAs remain unknown[5–8]. We hypothesized that the nuclear localization observed for some lncRNAs and mRNAs is encoded in short sequence elements that can autonomously dictate nuclear enrichment in an otherwise efficiently exported transcript. To systematically identify such regions, we designed a library of 5,511 sequences, each 109 nucleotides (nt) long, composed of fragments that tile the exonic sequences of 37 human lncRNAs, 13 3′ untranslated regions (UTRs) of mRNAs enriched in the nucleus in mouse liver[3], and 4 homologues of the abundant nuclear lncRNA *MALAT1* (Supplementary Tables 1, 2). These tiles were cloned into the 5′ and 3′ UTRs of *Aequorea coerulescens* (*Ac*)*GFP* mRNA (NucLibA library, Fig. 1a). Offsets between consecutive tiles were typically 25 nt (10 nt in *MALAT1* and *TERC* and 50 nt in the longer *XIST* and *MEG3* genes). After the constructs were transfected into MCF7 cells in triplicate, we sequenced the inserts from *GFP* mRNAs from whole cell extracts (WCE), nuclear and cytoplasmic fractions, and the input plasmids (at least ten million reads per sample; Supplementary Table 3). Normalized counts of unique molecular identifiers (UMIs; Supplementary Table 2) were used to evaluate the effect of each inserted tile on the expression level and subcellular localization of the *GFP* mRNA (Supplementary Table 4).

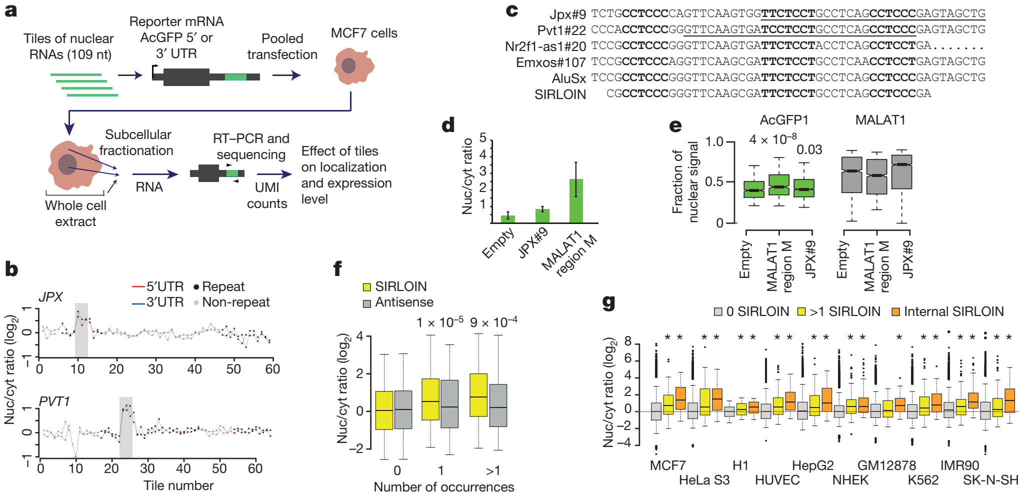


Figure 1 | NucLibA analysis and the SIRLOIN element. a, Experimental approach. UMI, unique molecule identifier. **b**, Nuclear/cytoplasmic (nuc/cyt) ratios of tiles cloned into the indicated region of *AcGFP*. Tiles overlapping repetitive elements are in black and other tiles are in grey. The region that overlaps the SIRLOIN element is shaded. **c**, Sequence alignment of four of the most effective tiles, the consensus sequence of the AluSx repeat family, and the SIRLOIN element. C/T-rich hexamers are in bold text, and the regions in JPX#9 and PVT1#22 that were mutagenized in NucLibB are underlined. **d**, Determination by qRT–PCR of the abundance of *AcGFP* mRNA with the indicated fragments cloned into the 3′ UTR. $n = 10–14$ independent fractionations. Mean ± s.e.m. is shown. **e**, Imaging flow cytometry of *AcGFP* and *MALAT1* mRNA; the fraction of the signal that overlaps with DAPI signal out of the total signal intensity is displayed. Values above boxes indicate *P* values compared to control (two-sided Wilcoxon test). $n = 570–3,621$ independent cells. Boxplots show the 5th, 25th, 50th, 75th and 95th percentiles. **f**, **g**, Nuc/cyt ratios for RNAs with the indicated number of SIRLOINs or their antisense in MCF7 cells (**f**) and each of ten ENCODE cell lines (**g**). Boxplots show the 5th, 25th, 50th, 75th and 95th percentiles. In **f**, numbers above boxes show *P* values for sense versus antisense comparisons (two-sided Wilcoxon). In **g**, $^*P < 0.01$ (two-sided Wilcoxon) when compared to transcripts without SIRLOINs. $n > 185$ genes in each group.

[1] Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel.

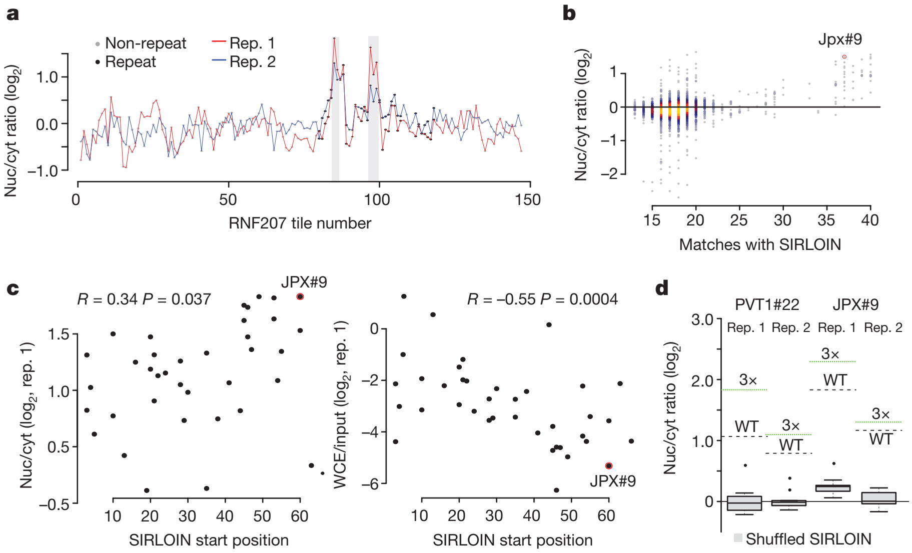


Figure 2 | NucLibB analysis. a, As in Fig. 1b, but showing effects of the tiles in *RNF207* mRNA. **b**, Similarity to a SIRLOIN element (number of matching bases, without allowing indels) and nuc/cyt ratios in NucLibB. Colouring indicates the local point density (yellow, high; grey, low). Only wild-type tiles were used; $n = 1{,}465$. **c**, SIRLOIN position in the tile and localization (left) or expression level (right). Spearman correlation, $n = 52$. **d**, Boxplots show 5th, 25th, 50th, 75th and 95th percentiles of nuc/cyt ratios of shuffled, dinucleotide-preserving sequences of the indicated tiles ($n = 10$). Horizontal lines show the effects of the wild-type (WT) sequence and of sequences containing three core parts of the SIRLOIN element (3×) (underlined in Fig. 1c).

To identify high-confidence effects, we focused on consecutive tiles that had a consistent effect on localization. We identified 19 regions from 14 genes spanning 2–4 overlapping tiles, with each of the tiles associated with over 30% nuclear enrichment (Supplementary Table 5). The three regions in which the tiles showed the highest enrichment originated from the lncRNAs *JPX*, *PVT1* and *NR2F1-AS1*, and the tiles had similar activity when placed in either the 3′ or 5′ UTRs of the *GFP* mRNA (Fig. 1b and Extended Data Fig. 1). These tiles overlapped Alu repeat sequences inserted in an 'antisense' orientation, and the overlap region between the active patches converged on a 42-nt fragment that contained three stretches of at least six pyrimidines (C/T), two of which were similar to each other and matched the consensus RCCTCCC (where R denotes A/G). We named this 42-nt sequence element SIRLOIN (SINE-derived nuclear RNA localization) (Fig. 1c). An individually cloned SIRLOIN-containing tile could drive nuclear enrichment of *GFP* mRNA, measured using quantitative PCR with reverse transcription (qRT–PCR) (Fig. 1d; 'region M' of *MALAT1* (about 600 nt)[5] is used as a positive control) and imaging flow cytometry with the PrimeFlow RNA assay (Fig. 1e).

Notably, although there was no significant correlation between the effects of individual tiles on expression levels and on localization ($R = 0.001$, $P = 0.82$), SIRLOIN-containing tiles were associated with consistently lower *GFP* RNA expression levels ($R = -0.31$ between effects on localization and expression, $P = 8.5 \times 10^{-3}$). SIRLOIN elements thus affect both the localization and the expression level of transcripts.

In RNA sequencing (RNA-seq) data from subcellular fractions[2], SIRLOIN elements were associated with nuclear enrichment in MCF7 cells (Fig. 1f and Extended Data Fig. 2a), and two or more SIRLOIN elements were associated with significant ($P < 0.01$, Wilcoxon test; 1.24–2.41-fold on average) nuclear enrichment in nine out of ten ENCODE cell lines (Fig. 1g). Notably, single SIRLOIN elements in internal exons were associated with stronger nuclear enrichment than those in terminal exons (Fig. 1g, 1.35–2.84-fold change on average), with similar trends observed for both mRNAs and lncRNAs (Extended Data Fig. 2b, c). When a gene had well-expressed alternative isoforms with and without SIRLOIN elements in internal exons, those isoforms with the SIRLOIN element were more nuclear (1.07–1.53-fold on average, $P < 0.05$ in seven out of ten cell lines, Wilcoxon test; Extended Data Fig. 3a and Supplementary Note 1). When we compared nucleoplasmic, chromatin and nucleolar fractions in K562 cells, SIRLOIN-overlapping transcripts were significantly depleted from the chromatin and nucleolar fractions (Extended Data Fig. 3b), suggesting that those transcripts accumulate in the nucleoplasm. There was no significant difference in mRNA half-lives between SIRLOIN-overlapping and other mRNAs[9] (Extended Data Fig. 3c).

Alu elements are the most common SINE elements in the human genome, covering about 10% of the genome sequence[10]. Alu elements are enriched in transcribed regions and have had a substantial effect on transcriptome evolution, for example through the contribution of new exons[11] and polyadenylation sites[12,13]. Such events are actively suppressed in mRNAs[14], but are common in lncRNAs[15]. Alu elements have also been reported to act as functional modules in lncRNAs via intramolecular[7] and intermolecular[16] pairing with other Alus[17]. SIRLOIN elements are quite common—13.1% of lncRNAs and 7.5% of human mRNAs have a SIRLOIN element, and 3.4% and 0.3%, respectively, have a SIRLOIN element in an internal exon, which we find to be more effective at inducing nuclear localization. Exonization of Alu elements thus contributes to the tendency of lncRNAs to be enriched in the nucleus and expressed at lower levels than mRNAs.

We next designed and cloned into the 3′ UTR of the *AcGFP* mRNA a second library, NucLibB (Supplementary Tables 6–8), that included tiles from additional lncRNAs and mRNAs that overlapped SIRLOIN elements, and a large number of sequence variations within two 30-nt fragments of two of the most effective tiles in NucLibA (JPX#9 and PVT1#22). Using NucLibB we validated the nuclear-enrichment activity of the *JPX* and *PVT1* SIRLOINs and identified 18 additional SIRLOIN-overlapping regions from five lncRNAs and five mRNAs that led to nuclear enrichment (Supplementary Table 9). In some cases, several regions from the same transcript were effective in mediating nuclear localization (Fig. 2a and Extended Data Fig. 4) and SIRLOIN-matching tiles were consistently active (Fig. 2b). SIRLOIN-containing sequences in NucLibB also affected expression levels, with a strong negative correlation between effects on expression and nuclear enrichment (Spearman $R = -0.6$, $P < 10^{-16}$; Extended Data Fig. 5a). Notably, among the SIRLOIN-containing sequences in NucLibB, those that had a SIRLOIN closer to the 3′ end were more enriched in the nucleus ($R = 0.34$, $P = 0.037$) and more poorly expressed ($R = -0.55$,

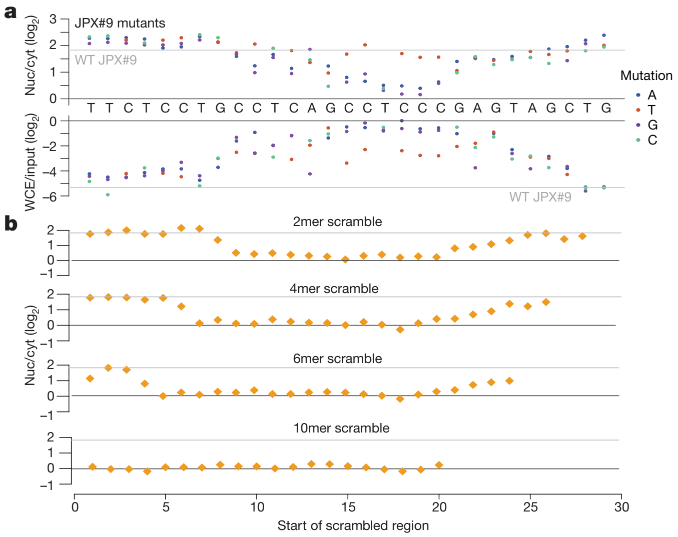


Figure 3 | Effect of SIRLOIN sequence changes on element activity. **a**, Localization (top) and expression level (bottom) of sequence variants that are identical to JPX#9 except for the indicated base change. Horizontal grey lines indicate the corresponding level for the wild-type sequence. **b**, Localization of JPX#9 sequences in which the indicated number of bases was scrambled with C↔G and A↔T changes. The position of the point indicates the first base of the scrambled region.

$P = 0.0004$), suggesting that the context of SIRLOINs affects their activity (Fig. 2c).

Shuffled SIRLOIN sequences of JPX#9 and PVT1#22 (preserving their dinucleotide composition) did not impose nuclear enrichment, as expected (Fig. 2d). Sequences with three repeats of core SIRLOIN parts from JPX#9 and PVT1#22 caused stronger nuclear enrichment than sequences containing only one SIRLOIN element (3.26-fold versus 2.4-fold on average; Fig. 2d). NucLibB also contained sequences composed of repetitions of individual 6- or 10-mers from the cores of JPX#9 and PVT1#22 SIRLOIN elements, separated by AT dinucleotides. Although we observed more nuclear-enrichment activity from the C/T-rich elements compared to other elements (Extended Data Fig. 5b, c), such sequences exerted very limited effects on either nuclear enrichment or expression level, suggesting that secondary structure or sequences beyond the C/T-rich motifs are important for SIRLOIN function.

Single point mutations to purines (A/G) in the second RCCTCCC motif of JPX#9 were sufficient to abolish the effect of the sequence on both localization and expression level, whereas C→T mutations in that region had little effect (Fig. 3a and Extended Data Fig. 5d). More extensive changes, alternating A↔T and G↔C, were also deleterious when made outside the RCCTCCC motif, suggesting that additional parts of SIRLOIN are essential for its function but can tolerate single base changes (Fig. 3b). The PVT1#22 sequence was more resilient to changes than the JPX#9 sequence, and single point mutations had a limited effect on its activity (Extended Data Fig. 5e). More extensive changes to PVT1#22, including mutating four bases in the 3′ part of its SIRLOIN element, were sufficient to abolish its activity (Extended Data Fig. 5f). We conclude that the second RCCTCCC motif is the most important part of the SIRLOIN element, but that other SIRLOIN regions also contribute to its function.

We next hypothesized that SIRLOIN elements act through interactions with specific RNA-binding proteins. To test this hypothesis, we queried ENCODE eCLIP datasets for protein–RNA interaction sites that were specifically enriched in SIRLOINs (see Methods). HNRNPK, an abundant nuclear RNA-binding protein with known roles in the biology of specific lncRNAs[18,19], ranked first among 112 factors in our eCLIP analysis (Supplementary Table 11 and Fig. 4a). Reassuringly, the motif enriched in the HNRNPK binding peaks throughout the transcriptome, as identified by GraphProt[20] (see Methods), was a pyrimidine-rich sequence with a CCTCC core (Fig. 4b), consistent with the known preferences of HNRNPK for C-rich sequences[21], and matching the RCCTCCC sequence we identified in our mutagenesis analysis. Moreover, the three KH RNA-binding domains of HNRNPK have been shown to act cooperatively in binding sequences with triplets of C/T-rich regions[22], fitting the sequence architecture of the SIRLOIN elements. GraphProt analysis also suggested the C/T-rich motif is preferentially bound in a structured context (Fig. 4b), consistent with our observation that simple repeats of short motifs from SIRLOIN were not functional. Using RNA immunoprecipitation (RIP) with an HNRNPK-specific antibody, we confirmed that HNRNPK binds to *AcGFP* mRNA supplemented with SIRLOIN elements, but not to a single-nucleotide mutant that was not effective for nuclear enrichment (Fig. 4c). Tethering of an HNRNPK protein to the 3′ UTR of a luciferase mRNA using a λN peptide-BoxB system[23] was sufficient to induce roughly threefold nuclear enrichment, suggesting a direct causal role of HNRNPK in the process (Fig. 4d).

We then tested whether HNRNPK binding was also associated with nuclear enrichment in transcripts that do not contain SIRLOIN elements. Notably, the number of HNRNPK binding clusters in eCLIP data correlated with stronger nuclear enrichment of lncRNAs and mRNAs in HepG2 cells (Fig. 4e; Spearman $R = 0.14$, $P < 10^{-16}$; see Supplementary Table 11 for correlations with other factors) and, to a lesser extent, in K562 cells ($R = 0.05$, $P = 3.6 \times 10^{-7}$; Extended Data Fig. 6a), and this correlation was essentially identical when we considered only transcripts that did not contain any SIRLOIN elements ($R = 0.15$ for HepG2 cells). The correlation between HNRNPK binding events and nuclear enrichment was highly significant when controlling for transcript length and expression level ($R = 0.2$, $P < 10^{-16}$ in HepG2 cells and $R = 0.07$, $P < 10^{-16}$ in K562 cells). Simple counts of C-rich motifs in transcript sequences were also significantly correlated with nuclear enrichment, with stronger correlations when considering counts only in the internal exons (Supplementary Note 2 and Extended Data Fig. 7). By contrast, in K562 cells there was no correlation between the number of HNRNPK eCLIP clusters and chromatin/nucleoplasm ratios ($R = -0.006$, $P = 0.57$; $R = -0.01$, $P = 0.0169$ when controlling for expression levels and transcript length). There was no correlation between nuclear enrichment and the number of eCLIP binding peaks of a different poly(C) binding protein, PCBP2 (HepG2 cells, Fig. 4e).

To validate the role of HNRNPK in regulation of its bound targets, we knocked down HNRNPK in MCF7 cells using short interfering RNAs (siRNAs; Extended Data Fig. 6b–d). HNRNPK knockdown had a substantial effect on subcellular enrichment of hundreds of genes, with 397 genes becoming twofold more nuclear and 283 genes becoming more cytoplasmic. The decrease in nuclear enrichment was significantly correlated with the number of HNRNPK eCLIP clusters (Spearman $R = -0.22$ between change in nuclear/cytoplasmic ratio and number of eCLIP clusters, $P < 10^{-16}$; $R = -0.26$ partial correlation controlling for transcript length and expression level in MCF7 cells; Fig. 4f); with similar effects observed in mRNAs and lncRNAs (Extended Data Fig. 6e). Genes with more than one SIRLOIN element were significantly less enriched in the nucleus following HNRNPK knockdown (Fig. 4f and Extended Data Fig. 6f). Notably, in eCLIP-defined targets, changes in nuclear enrichment were mostly due to a reduction in transcript levels in the nucleus (21% on average) accompanied by a mild increase in cytoplasmic levels (10% on average), overall resulting in slightly decreased expression levels of HNRNPK targets following knockdown (4% on average, Fig. 4g). Similar results were obtained following HNRNPK knockdown in HeLa cells (Extended Data Fig. 8), and when testing the NucLibB library following HNRNPK knockdown (Fig. 4h and Supplementary Note 3). These results suggest that HNRNPK binding drives nuclear enrichment mediated by SIRLOIN elements, but other factors are likely to contribute to the decrease in overall expression level associated with SIRLOIN integration.

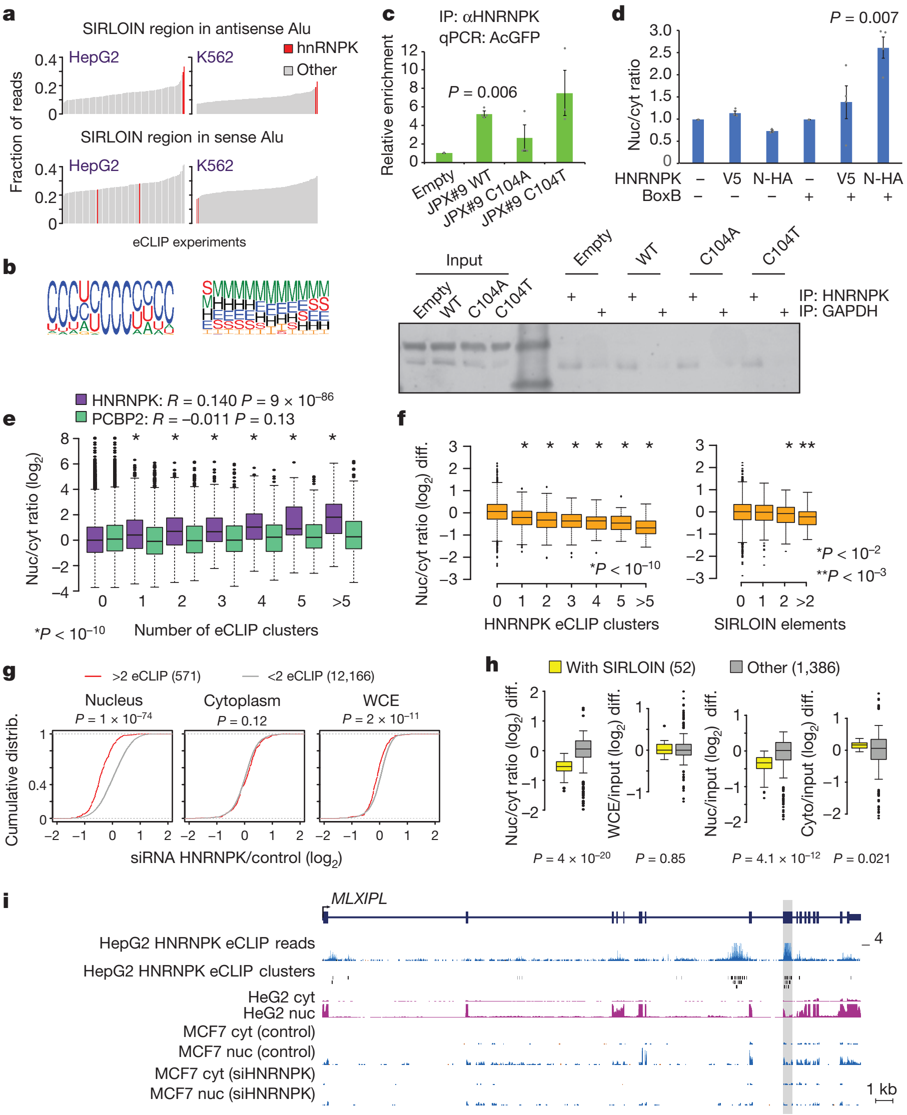


Figure 4 | HNRNPK drives nuclear localization of SIRLOIN-containing transcripts. a, HNRNPK binding sites are enriched in the SIRLOIN element within Alu repeats (see Methods). **b**, Sequence (left) and structure (right) logos of the motif enriched in HNRNPK eCLIP clusters (HepG2 cells, first replicate), as identified by GraphProt. Structural annotation: stems (S), external regions (E), hairpins (H), internal loops (I), multiloops (M) and bulges (B). **c**, Enrichments of *AcGFP* RNA with the indicated inserts following immunoprecipitation (IP) using an HNRNPK antibody, normalized to the GAPDH antibody. $n = 4$ independent experiments. P values computed using two-sided t-test. Mean ± s.e.m. shown. **d**, Localization of Renilla luciferase mRNA with and without five BoxB hairpins in the 3′ UTR, either without expression of exogenous HNRNPK (−) or with expression of an HNRNPK fused to the indicated tag (V5 or N-HA, which binds BoxB). $n = 3$–4 independent experiments; P values computed using two-sided t-test. Bars show mean ± s.e.m. **e**, Localization versus number of eCLIP clusters in HepG2 cells. Asterisks indicate significant difference between genes with the indicated number of clusters and genes without clusters (two-sided Wilcoxon). Boxplots show 5th, 25th, 50th, 75th and 95th percentiles. $n > 60$ genes in each group. **f**, Change in the nuc/cyt ratio following HNRNPK knockdown. P values computed using two-sided Wilcoxon compared to genes with no eCLIP clusters or no SIRLOIN elements. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles. $n > 60$ genes in each group. **g**, Changes in expression following HNRNPK knockdown in the indicated sample (two-sided Wilcoxon). **h**, Changes in the indicated ratios for NucLibB fragments following HNRNPK knockdown. Each plot shows the difference between the ratio following transfection of HNRNPK siRNA and a non-targeting control. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles. P values computed using two-sided Wilcoxon test. **i**, HNRNPK binding and the *MLXIPL* gene. eCLIP reads and clusters are from HepG2 cells (replicate 1) from the ENCODE data portal. Expression levels in HepG2 nuclear fraction were capped to allow visual comparison. The exon enriched with HNRNPK binding is shaded.

Our observation that repeated C/T-rich elements are globally associated with nuclear enrichment and lower expression level is also supported by previous studies of individual mRNAs and lncRNAs. For example, nuclear retention elements have previously been associated with decreased overall expression levels in *HBB* mRNA[24]. An AGCCC motif has been reported to be important for nuclear retention of the *BORG* lncRNA[6]. A survey of known nuclear enrichment elements in ExportAid[25] revealed several viral sequences that contained closely spaced C/T hexamers, such as the *cis*-acting inhibitory element (CIE) of human T-cell lymphotropic virus type 1 and the two short polypyrimidine tracts associated with nuclear retention in hepatitis B virus[26]. Viral RNAs may thus also rely on HNRNPK-mediated nuclear enrichment for maintaining low expression levels and nuclear retention during latency. Some well-studied nuclear lncRNAs, such as *XIST* and *NEAT1*, did not contain any regions that exhibited consistent nuclear-enrichment activity in our system, suggesting that the regions that encode nuclear enrichment in these lncRNAs are either longer than the 109-nt tiles or not active in our specific expression context. The roughly

600-nt region M sequence in *MALAT1*[5] causes strong nuclear enrichment in our system (Fig. 1e), and longer regions can drive nuclear localization in *XIST* and other lncRNAs[27], so we suggest that multiple independent pathways are likely to be responsible for nuclear enrichment in lncRNAs and mRNAs, recognizing specific RNA sequences and/or other features of the ribonucleoprotein. It is also likely that the nuclear retention of individual RNA molecules is affected by more than one pathway. As the contribution of each SIRLOIN element to nuclear enrichment, at least in the UTR context, is around twofold, additional pathways are likely to contribute to localization of those SIRLOIN-containing RNAs that are strongly enriched in the nucleus (for example, *PVT1*[28]).

The pathway uncovered here is likely to be relevant for at least some of the transcripts that have been observed to be enriched in the nucleus *in vivo* in mice[3] (Supplementary Note 4). For instance, the mRNA of the transcription factor *MLXIPL* (also known as ChREBP) is retained in nuclear speckles in the mouse liver, beta cells and intestine[3]. *MLXIPL* has a long internal exon containing multiple HNRNPK binding sites, is strongly enriched in the nucleus in various human cell lines and was strongly affected by HNRNPK knockdown in MCF7 cells (Fig. 4i). We also identified a SIRLOIN-like element in the mouse B1 repeats, and together with the human SIRLOIN element it appears to have contributed to divergence in expression level and localization between the two species (Supplementary Note 4 and Extended Data Fig. 9).

The mechanism by which HNRNPK affects nuclear enrichment is currently unclear. As HNRNPK interacts with splicing factors and can affect the splicing of specific genes[29], and intron retention is associated with nuclear retention[30], the HNRNPK-induced nuclear enrichment could be mediated by effects on RNA splicing; however, this is unlikely to be the case (Supplementary Note 5). RNA editing, which has been suggested to have a causal role in nuclear retention of inverted Alu repeats[7], is also unlikely to be involved (Supplementary Note 6 and Extended Data Fig. 10).

The nuclear enrichment mechanism we discovered is more common in lncRNAs, but also employed by some mRNAs. This highlights how studying lncRNAs, which use unique functional mechanisms and are under different selective pressures than mRNAs, can generally enhance our understanding of RNA biology. We expect that increasing understanding of the repertoire of lncRNA functions in cells will enable similar high-throughput approaches for identification of additional sequence and structural elements shared across lncRNAs, and expedite classification of these enigmatic genes into functional families.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 May; accepted 29 December 2017.
Published online 24 January 2018.

1. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
2. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
3. Bahar Halpern, K.*et al.* Nuclear retention of mRNA in mammalian tissues. *Cell Reports* **13,** 2653–2662 (2015).
4. Battich, N., Stoeger, T. & Pelkmans, L. Control of transcript variability in single mammalian cells. *Cell* **163,** 1596–1610 (2015).
5. Miyagawa, R. *et al.* Identification of *cis*- and *trans*-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. *RNA* **18,** 738–751 (2012).
6. Zhang, B. *et al.* A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Mol. Cell. Biol.* **34,** 2318–2329 (2014).
7. Chen, L. L., DeCerbo, J. N. & Carmichael, G. G. Alu element-mediated gene silencing. *EMBO J.* **27,** 1694–1705 (2008).
8. Prasanth, K. V. *et al.* Regulating gene expression through RNA nuclear retention. *Cell* **123,** 249–263 (2005).
9. Schueler, M. *et al.* Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* **15,** R15 (2014).
10. Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13,** 1998–2004 (2003).
11. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* **300,** 1288–1291 (2003).
12. Chen, C., Ara, T. & Gautheret, D. Using Alu elements as polyadenylation sites: A case of retroposon exaptation. *Mol. Biol. Evol.* **26,** 327–334 (2009).
13. Tajnik, M. *et al.* Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends. *Nucleic Acids Res.* **43,** 10492–10505 (2015).
14. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152,** 453–466 (2013).
15. Kelley, D. R. & Rinn, J. L. Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol.* **13,** R107 (2012).
16. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470,** 284–288 (2011).
17. Johnson, R. & Guigó, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20,** 959–976 (2014).
18. Dimitrova, N. *et al. LincRNA-p21* activates *p21* in *cis* to promote Polycomb target gene expression and to enforce the G1/S checkpoint. *Mol. Cell* **54,** 777–790 (2014).
19. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161,** 404–416 (2015).
20. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15,** R17 (2014).
21. Choi, H. S. *et al.* Poly(C)-binding proteins as transcriptional regulators of gene expression. *Biochem. Biophys. Res. Commun.* **380,** 431–436 (2009).
22. Paziewska, A., Wyrwicz, L. S., Bujnicki, J. M., Bomsztyk, K. & Ostrowski, J. Cooperative binding of the hnRNP K three KH domains to mRNA targets. *FEBS Lett.* **577,** 134–140 (2004).
23. Baron-Benhamou, J., Gehring, N. H., Kulozik, A. E. & Hentze, M. W. Using the λN peptide to tether proteins to RNAs. *Methods Mol. Biol.* **257,** 135–154 (2004).
24. Akef, A., Lee, E. S. & Palazzo, A. F. Splicing promotes the nuclear export of β-globin mRNA by overcoming nuclear retention elements. *RNA* **21,** 1908–1920 (2015).
25. Giulietti, M., Milantoni, S. A., Armeni, T., Principato, G. & Piva, F. ExportAid: database of RNA elements regulating nuclear RNA export in mammals. *Bioinformatics* **31,** 246–251 (2015).
26. Roy, D., Bhanja Chowdhury, J. & Ghosh, S. Polypyrimidine tract binding protein (PTB) associates with intronic and exonic domains to squelch nuclear export of unspliced RNA. *FEBS Lett.* **587,** 3802–3807 (2013).
27. Shukla, C. J. *et al.* High-throughput identification of RNA nuclear enrichment sequences. *EMBO J.* e98452 (2018).
28. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16,** 20 (2015).
29. Bomsztyk, K., Denisenko, O. & Ostrowski, J. hnRNP K: one protein multiple processes. *BioEssays* **26,** 629–638 (2004).
30. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24,** 1774–1786 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Pillai, S. Schwarz, S. Itzkovitz, N. Stern Ginossar, and members of the Ulitsky laboratory for discussions and comments on the manuscript, and Z. Porat from the Weizmann Institute FACS core for assistance with Imaging Flow Cytometry. This research was supported by the Israeli Centers for Research Excellence (1796/12); Israel Science Foundation (1242/14 and 1984/14); European Research Council lincSAFARI; Minerva Foundation; Lapon Raymond; and the Abramson Family Center for Young Scientists. I.U. is incumbent of the Sygnet Career Development Chair for Bioinformatics.

Author Contributions Y.L. and I.U. conceived and designed the study. Y.L. carried out all experiments and I.U. carried out computational analysis. Y.L. and I.U. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to I.U. (igor.ulitsky@weizmann.ac.il).

Reviewer Information *Nature* thanks J. Ule, C. Wahlestedt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cell culture and transfection. MCF7 and HeLa cells (ATCC) were grown in DMEM (Gibco) containing 10% fetal bovine serum and penicillin/streptomycin mixture (1%) at 37 °C in a humidified incubator with 5% CO_2. Cell lines have not been authenticated and were routinely tested for mycoplasma contamination. Plasmid transfections were performed using PolyEthylene Imine (PEI)[31] (PEI linear, M_r 25,000, Polyscience). RNA was extracted 24 h after library transfections.

Library design. Oligonucleotide pools were purchased from Twist Bioscience. Tiles overlapping EcoRI, BglII and BamHI recognition sequences were excluded from both libraries and tiles overlapping HindIII, XbaI and NotI were also excluded from NucLibA.

Plasmid library construction. The oligonucleotide pool was amplified by PCR (20 ng template in 4.8 ml reaction, divided into 96 50-μl reactions; see Supplementary Table 12 for primer sequences), concentrated using Amicon tubes (UFC503096, Millipore) and purified using AMpure beads (A63881, Beckman) at a 2:1 beads:sample ratio according to the manufacturer's protocol. The insert was digested with EcoRI and BglII and cloned into likewise-digested pAcGFP1-C1 or pAcGFP1-N1 (Clontech) for 3′ and 5′ insertion, respectively. The ligation was transformed into *Escherichia coli* electrocompetent bacteria (60117-2, Lucigen) and plated on 15 × 15-cm LB/Amp agar plates. Colonies were scraped off the plates and DNA was extracted using a plasmid maxi kit (12163, Qiagen).

Extraction of cytoplasmic and nuclear RNA. Cells were washed twice in cold PBS and resuspended in 300 μl RLN buffer (50mM Tris•Cl pH 8, 140 mM NaCl, 1.5 mM $MgCl_2$, 10 mM EDTA, 1 mM DTT, 0.5% NP-40, 10 U/ml RNase inhibitor) and incubated on ice for 5 min. The extract was centrifuged for 5 min at 300*g* in a cold centrifuge and the supernatant was transferred to a new tube and centrifuged again for 1 min at 500*g* in a cold centrifuge; the supernatant (cytoplasmic fraction) was transferred again to a new tube and RNA was extracted using TRIREAGENT (MRC). The nuclear pellet was washed once in 300 μl RLN buffer and resuspended in 1 ml buffer S1 (250 mM sucrose, 10 mM $MgCl_2$, 10 U/ml RNase inhibitor), layered over 3 ml buffer S3 (880 mM sucrose, 0.5 mM $MgCl_2$, 10 U/ml RNase inhibitor), and centrifuged for 10 min at 2,800*g* in a cold centrifuge. The supernatant was removed and RNA was extracted from the nuclear pellet using TRIREAGENT.

Sequencing library generation. One microgram of RNA was used for cDNA production using the qScript Flex cDNA synthesis kit (95049, Quanta) and a gene-specific primer containing part of the Illumina RD2 region. The entire cDNA reaction was diluted into 100 μl second strand reaction with a mix of six primers introducing a unique molecular identifier (UMI) and a shift as well as part of the Illumina RD1 region. The second strand reaction was carried for a single cycle using Phusion Hot Start Flex DNA Polymerase (NEB, M0535), purified using AMpure beads at a 1.5:1 beads:sample ratio and eluted in 20 μl ddH_2O. Fifteen microlitres of the second strand reaction was used for amplification with barcoded primers, and the amplified libraries were purified by two-sided AMpure purification first with a 0.6:1 beads:sample ratio followed by a 1:1 ratio.

NucLibA samples were sequenced with 119-nt reads and the NucLibB samples with 75-nt paired-end reads on an Illumina NextSeq 500 machine.

Library data analysis. The sequenced reads were used to count individual library tiles using a custom Java script. We considered only R1 reads that contained the TTGATTCGATATCCGCATGCTAGC adaptor sequence, and extracted the unique molecular identifier (UMI) sequence preceding the adaptor. In R2 read, we removed the CGGCTTGCGGCCGCACTAGT adaptor and added the three bases preceding it to the UMI. Each read was then matched to the sequences in the library, without allowing indels. The matching allowed mismatches only at positions with Illumina sequencing quality of at least 35 and we allowed up to two mismatches in the first 15 nt ('seed'), and no more than four overall mismatches. If a read matched more than one library sequence, the sequence with the fewest mismatches was selected, and if the read matched more than one library sequence with the same number of mismatches, it was discarded. Per-read mismatches were counted for the RNA editing analysis. See Supplementary Table 3 for read mapping statistics. The output from this step was a table of counts of reads mapping to each library sequence in each library (Supplementary Tables 2, 7).

Only fragments with at least 10 reads on average in the WCE samples were used in subsequent analysis (5,153 fragments in NucLibA), and the number of UMIs mapping to each fragment was normalized to compute UPMs (UMIs per million sequenced UMIs). We then used these to compute nuclear/cytoplasmic and WCE/input ratios after adding a pseudocount of 0.5 to each UPM (Supplementary Tables 4, 8).

Human transcriptome analysis. The RefSeq transcript database (downloaded from the UCSC genome browser hg19 assembly on 30 June 2016) was used for all analyses, unless noted otherwise. Only transcripts with exonic lengths of at least 200 nt were considered, and for entries mapping to multiple genomic loci, only one of the loci was used. We quantified the isoform-level expression levels using RSEM[32] v1.2.31 in the ENCODE data (parameters –no-bam-output –bowtie2 –p 32 –forward-prob 0) and computed the average fraction of each isoform among the isoforms of each gene. Only isoforms with a relative abundance of more than 25% were considered. Fold-changes between nuclear and cytoplasmic fractions were computed using DESeq2[33] v1.12.4 with default parameters. A transcript was considered to contain a SIRLOIN element or its antisense if it aligned with the sequence (without indels) with no more than eight mismatches.

Alternative isoform analysis. For comparison of alternative isoforms containing SIRLOIN elements, we used the GENCODE v.26 transcript annotations, excluding transcripts with 'retained_intron' biotype to avoid likely targets of nonsense-mediated decay. Transcript expression levels were quantified using RSEM, and only transcripts with fragments per kilobase of transcript per million mapped reads (FPKM(nucleus) + FPKM(cytoplasm)) ≥ 1 on average across the two replicates and isoform usage >5% in at least one fraction were considered. For each of these transcripts, we computed the $\log_2$-transformed fold changes between the nuclear and cytoplasmic expression levels, adding a pseudocount of 0.5 to each FPKM and averaged the ratios between the two replicates. We then considered genes that had at least one isoform with a SIRLOIN in an internal exon and one isoform without, and averaged the nuclear/cytoplasmic ratios for the isoforms in each group, resulting in two values per gene. These were then compared using the Wilcoxon rank-sum test for paired samples. Selected pairs were then further validated by RT–PCR (Extended Data Fig. 3a). Cytoplasmic and nuclear RNA were extracted as described above and cDNA was generated with qScript Flex cDNA synthesis kit (95049, Quanta), using oligo-dT. PCR was performed using Phusion Hot Start Flex DNA Polymerase (NEB M0535).

eCLIP data analysis. For alignment of eCLIP reads to Alu elements, we built a STAR aligner[34] index using the *JPX* and *PVT1* Alu fragments, and aligned library reads to it using STAR with parameters:–outSAMstrandField intronMotif –readFilesCommand zcat –runThreadN 32 –outSAMtype BAM SortedByCoordinate –outWigType bedGraph read1_5p. We then computed for each experiment the number of R2 reads whose 5′ base mapped within the SIRLOIN fragment in either *PVT1* or *JPX* sequence compared to the rest of the Alu sequence in each orientation (Supplementary Table 10).

In order to identify the sequence and structural preferences of the motif in HNRNPK eCLIP clusters, we downloaded the eCLIP binding clusters identified by the ENCODE pipeline (accession ENCFF861DAK) from the ENCODE portal and used GraphProt[15] with default parameters ('train' module) to compare the eCLIP peaks to control peaks randomly sampled peaks of the same length from the same transcripts. The model identified by GraphProt had accuracy of 93% on this training and the sequence and structure motif logos were obtained from it using GraphProt with default parameters ('motif' module).

RNA immunoprecipitation. RIP was performed according to the native RIP protocol as described[35]. Anti-HNRNPK (RN019P, MLB) and anti-GAPDH (C-2118S, Cell Signaling) were pre-incubated for 1 h with protein-A/G magnetic beads (L00277, A2S) at 5 μg antibody to 50 μl bead slurry. Extract from 2 × 10^6 cells was added to the beads and incubated overnight with rotating at 4 °C. Precipitated RNA was extracted and analysed by qRT–PCR (see Supplementary Table 12 for primers).

Imaging flow cytometry. *MALAT1* and *AcGFP* mRNAs were labelled using the PrimeFlow RNA Assay kit (88-18009-204 Affymetrix) according to the manufacturer's protocol. *MALAT1* probes were labelled with Alexa Fluor 647 (VA1-11317) and *AcGFP* probes were labelled with Alexa Fluor 750 (VF6-14335); nuclei were stained with DAPI. Cells were visualized using the ImageStream^{X} Mark II (Amnis) and images were analysed using image analysis software (IDEAS 6.2; Amnis Corp). Images were compensated for fluorescent dye overlap by using single-stain controls. Cells were gated for single cells, using the area and aspect ratio features, and for focused cells, using the Gradient RMS feature, as previously described[36]. To calculate the nuclear fraction of each transcript, the intensity of either Alexa Fluor 647 or Alexa Fluor 750 within the nucleus (as defined by the morphology mask on the DAPI staining; channel 7) was measured, and divided by the total intensity for each cell.

RNA-seq. MCF7 or HeLa cells were transfected with 10 nM siRNA pool targeting HNRNPK or with control pool (Dharmacon) using Lipofectamin 3000 reagent (L3000001, Thermo Fisher). RNA was extracted 72 h after transfection and libraries were prepared using the SENSE mRNA-Seq Library prep kit (SKU: 001.24, Lexogen) according to the manufacturer's protocol.

RNA-seq reads were mapped to the human genome (hg19 assembly) with STAR[34] v020201 and transcript levels were quantified using RSEM[32] with parameters –no-bam-output –bowtie2 –p 32. Fold-changes were computed using DESeq2[39] with default parameters.

HNRNPK knockdown in cells expressing NucLibB. MCF7 cells were transfected with siRNA pool or non-targeting control as described above. Forty-eight hours after transfection, the cells were washed and transfected with NucLibB plasmid pool. RNA was extracted after an additional 24 h and libraries were prepared as described above. Libraries were analysed in the same way as described above and nuclear/cytoplasmic, nuclear/input, cytoplasmic/input and WCE/input ratios were used after adding a pseudocount of 0.5 to each UPM. Ratios for the two HNRNPK siRNA replicates were averaged.

Splicing analysis. MCF7 knockdown data was analysed using MISO[37] v.0.5.4 with default parameters comparing the HNRNPK knockdown samples to controls using differential splicing annotations obtained from the MISO website (miso_annotations_hg19_v2). After filtering for significant events (–num-inc 1 –num-exc 1 –num-sum-inc-exc 10 –delta-psi 0.20 –bayes-factor 10), only 45 events were significant in 30 comparisons (2 replicates, 3 fractions, 5 alternative splicing event types). Analysis of overall splicing efficiency was performed as previously described[38], evaluating the fraction of intron spanning reads out of all the reads that overlap splice junctions.

Conservation analysis. RefSeq transcripts for which we quantified nuclear/cytoplasmic ratios were mapped to Ensembl transcripts and human and mouse orthologues were obtained from Ensembl Compara 80. Nuclear/cytoplasmic ratios were compared between HepG2 cells (ENCODE data) and mouse liver[3] and expression levels between human liver expression (HPA data[39]) and ENCODE mouse liver expression, both quantified using RSEM (parameters –no-bam-output –bowtie2 –p 32). Only genes with FPKM $\geq$ 0.5 in at least one of the datasets were considered for the analysis. Overlaps with Alu and B1 elements were computed using RepeatMasker data from the UCSC genome browser.

Hexamer analysis. Sequences for the RefSeq transcripts described above were extracted from the UCSC genome browser. Occurrences of all possible hexamers (allowing overlaps) were counted either in all the exons, or just in the internal or terminal exons. Then, for each hexamer, the Spearman correlation coefficient was computed between the total number of occurrences of the motif and the nuclear/cytoplasmic ratio as computed by DESeq2 (with default parameters).

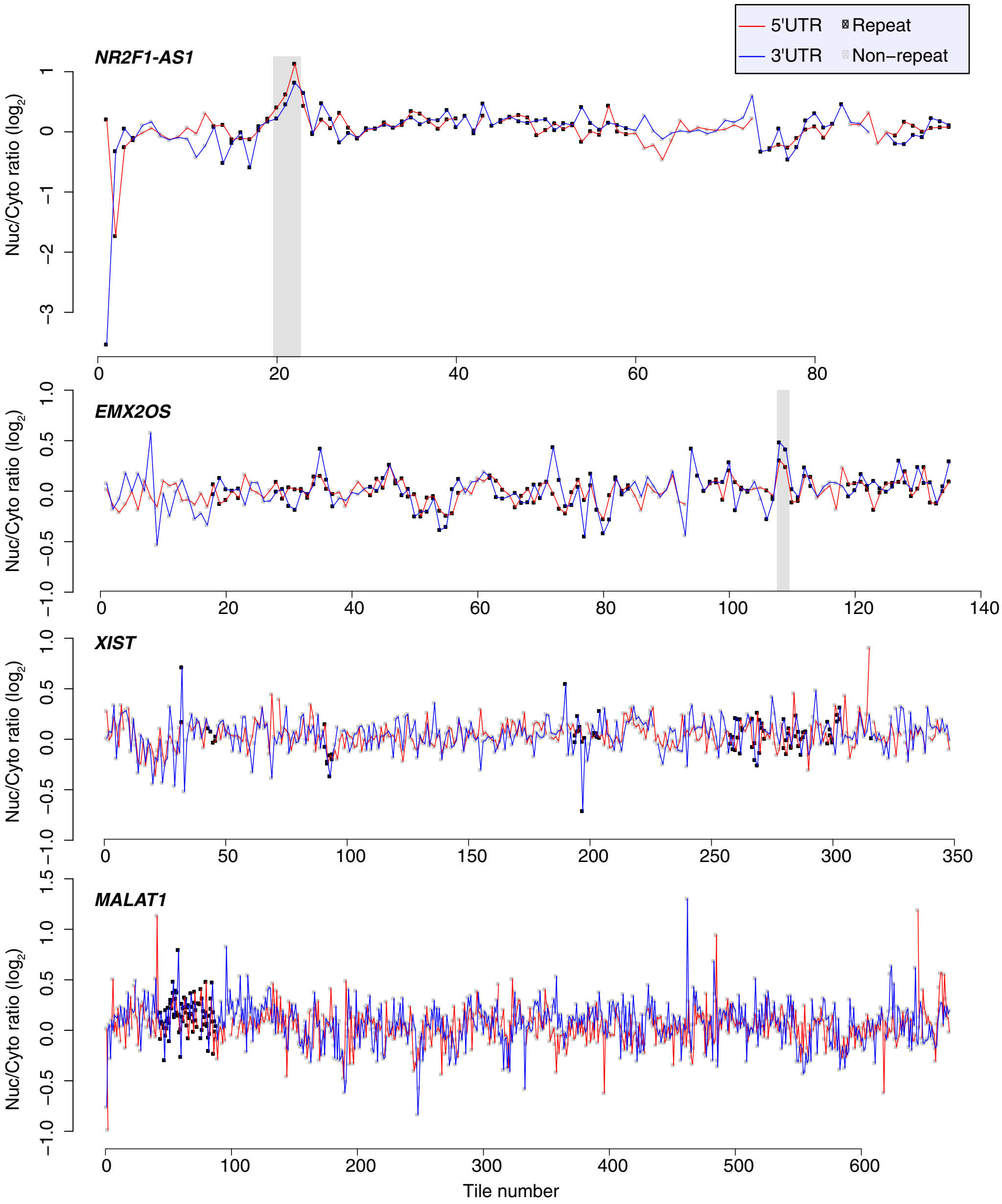
Tethering of hnRNPK to a reporter plasmid. pCIneo-RL-5×BoxB and pCIneo-N-HA were a gift from R. Pillai. pT7-V5-SBP-C1-HshnRNPK was a gift from E. Izaurralde (Addgene plasmid #64923). pCIneo-RL was generated by digestion of pCIneo-RL-5×BoxB with XbaI and XhoI; the overhangs were filled in by Klenow polymerase and ligated.

N-HA-hnRNPK vector was generated by swapping the V5-SBP in pT7-V5-SBP-C1-HshnRNPK with the N-HA region from pCIneo-N-HA. Cells were transfected with the reporters in the presence of the different hnRNPK constructs or of a control vector (pcDNA3.1), and cytoplasmic and nuclear RNA was extracted 48 h after transfection. The expression of the construct was verified by western blotting using anti-V5 antibody (Abcam ab206564) and anti-HA antibody (BioLegend HA.11).

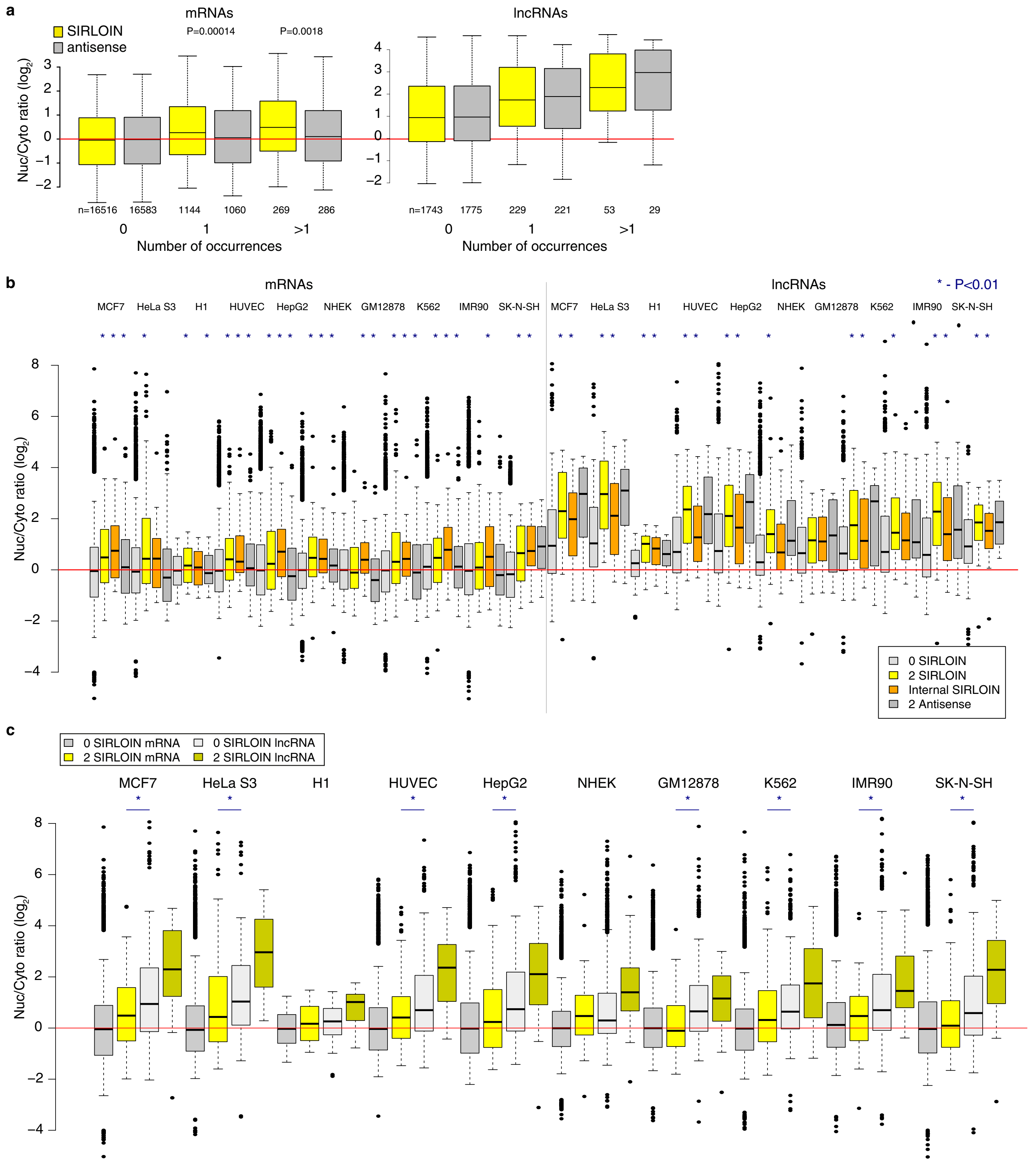
Statistics. All correlations were computed using Spearman correlation, unless otherwise indicated.

Data availability. All sequencing data are available in the SRA database, accession SRP111756. Source data for Figs 1b, d–g, 2a–d, 3a, b, and 4a, c–h are provided with the online version of the paper.

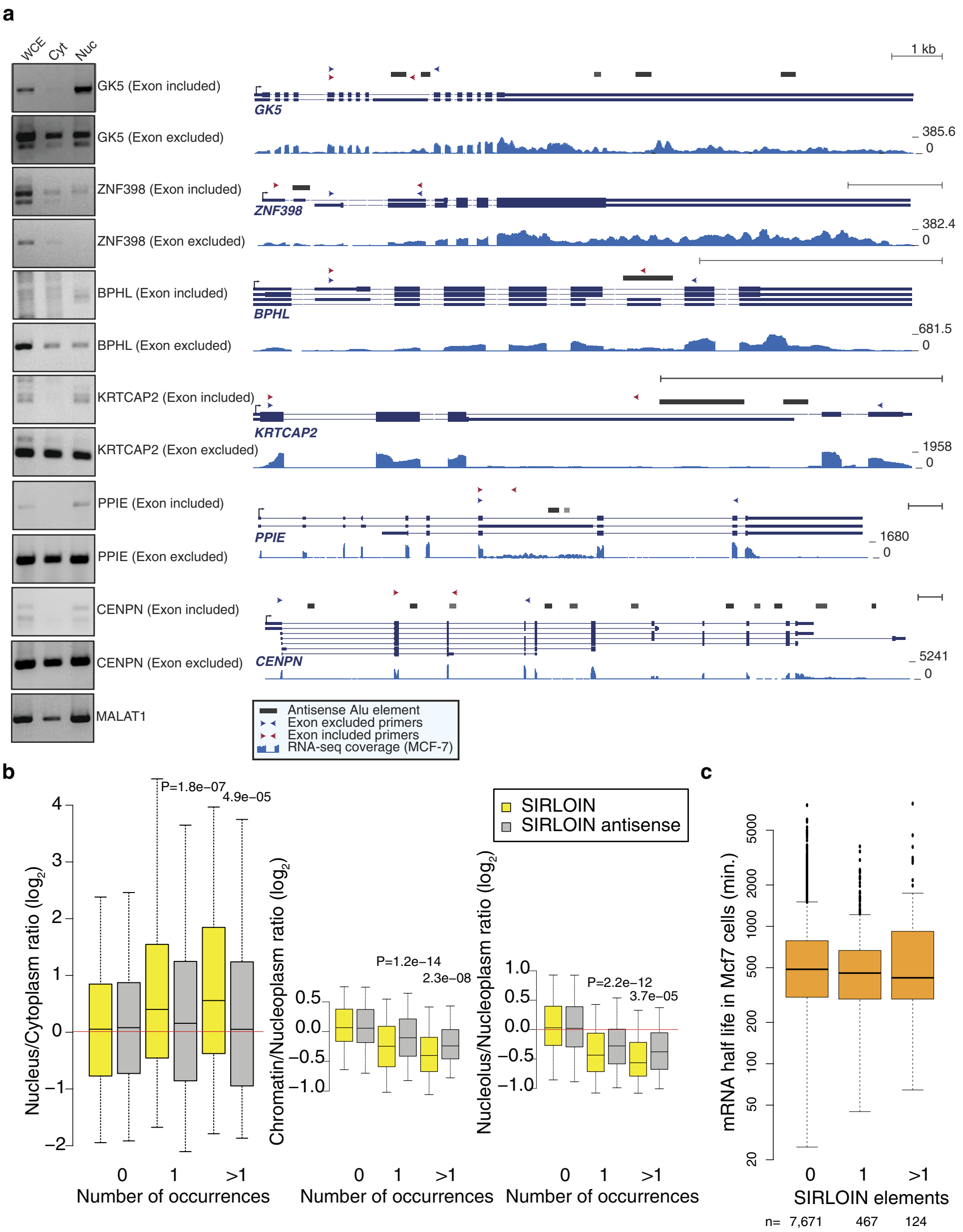
31. Durocher, Y., Perret, S. & Kamen, A. High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res.* **30,** E9 (2002).
32. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
35. Gagliardi, M. & Matarazzo, M. R. RIP: RNA immunoprecipitation. *Methods Mol. Biol.* **1480,** 73–86 (2016).
36. George, T. C. *et al.* Quantitative measurement of nuclear translocation events using similarity analysis of multispectral cellular images obtained in flow. *J. Immunol. Methods* **311,** 117–129 (2006).
37. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
38. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22,** 1616–1625 (2012).
39. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13,** 397–406 (2014).



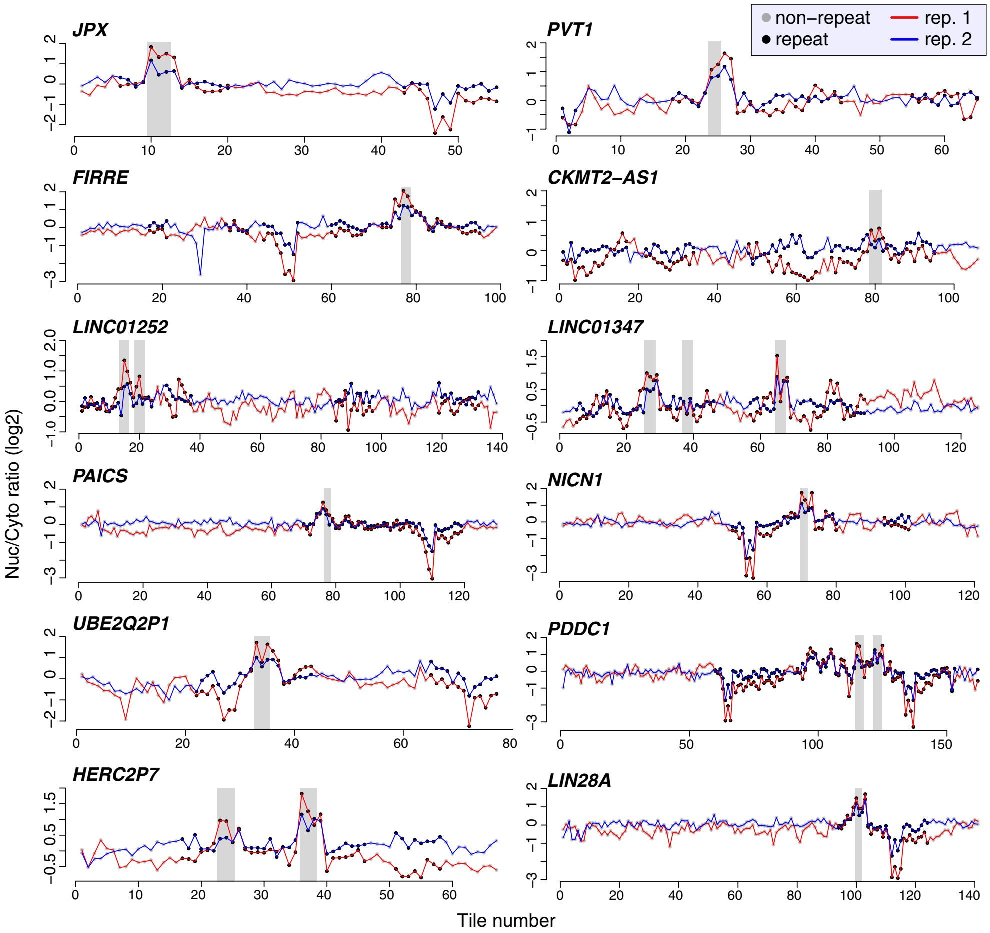
Extended Data Figure 1 | Effects of tiles in NucLibA on localization. Nuclear/cytoplasmic ratios for all the tiles in the indicated lncRNAs when cloned into the indicated region of *AcGFP*. Tiles overlapping repetitive elements are in black and other tiles are in grey. Regions with segments containing the SIRLOIN elements are shaded.



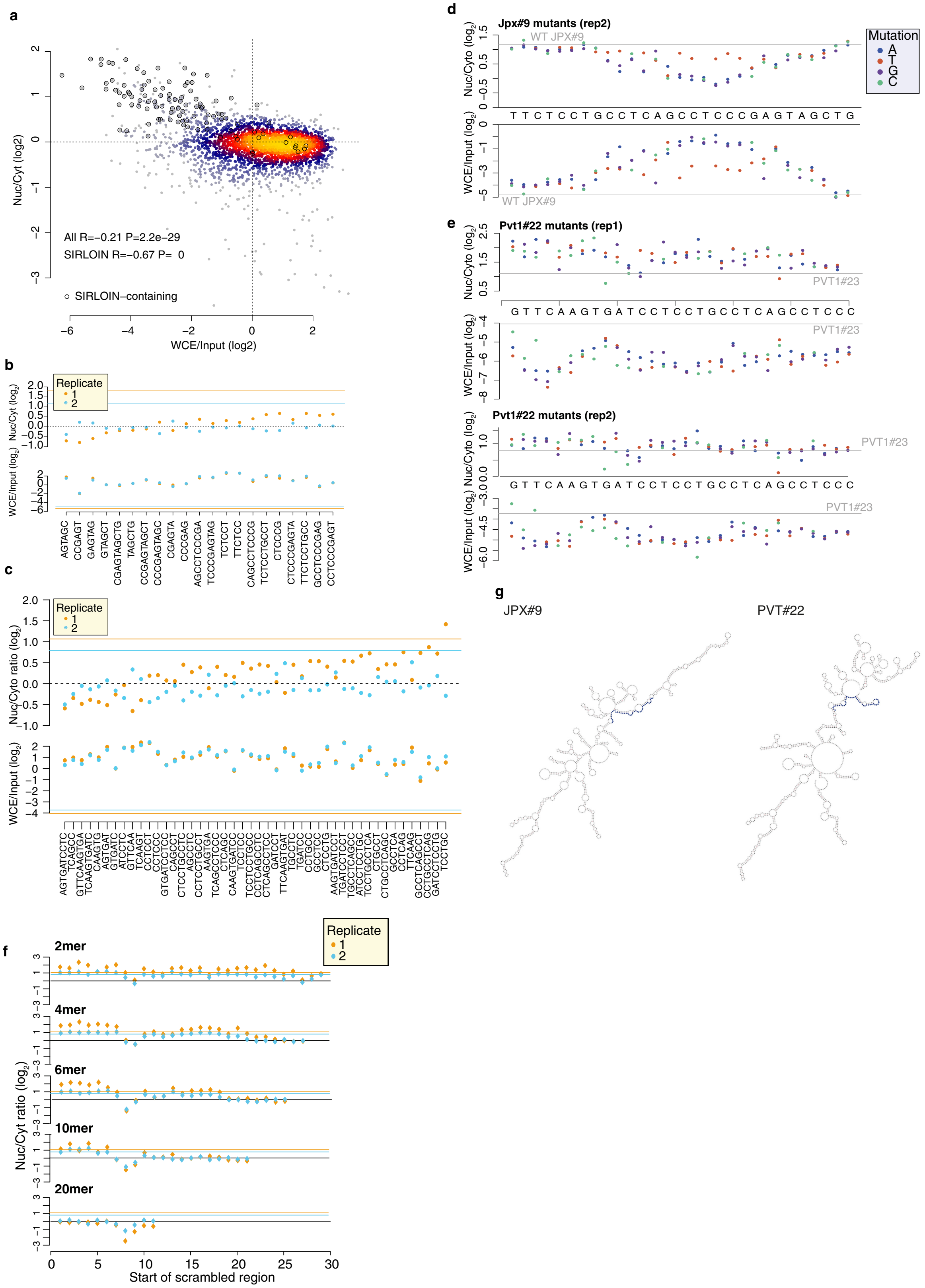
Extended Data Figure 2 | Nuclear/cytoplasmic ratios for lncRNAs and mRNAs. a, Nuclear/cytoplasmic expression ratios for mRNAs and lncRNAs containing the indicated number of SIRLOIN elements or SIRLOIN reverse complement (antisense) in MCF7 cells (ENCODE data). Otherwise, as in Fig. 1f. **b**, Nuclear/cytoplasmic expression ratio in mRNAs (left) and lncRNAs (right) in ten ENCODE cell lines. Transcripts without SIRLOIN elements, with two SIRLOIN elements, with a SIRLOIN element in one of the internal exons, or with two antisense SIRLOIN elements are compared. Asterisks above the '2 SIRLOIN' and 'Internal SIRLOIN' boxes indicate $P < 0.01$ by two-sided Wilcoxon test relative to the genes with no SIRLOIN elements. Asterisks above the '2 Antisense' boxes indicate $P < 0.01$ relative to the '2 SIRLOIN' boxes. Otherwise, as indicated in Fig. 1g. $n = 29$–16,694 genes per group. **c**, A rearranged version of part of **b** that facilitates visual comparison between mRNAs with at least two SIRLOIN elements and lncRNAs without SIRLOIN elements. Asterisks indicate $P < 0.01$ by two-sided Wilcoxon test. $n = 50$–16,694 genes per group.



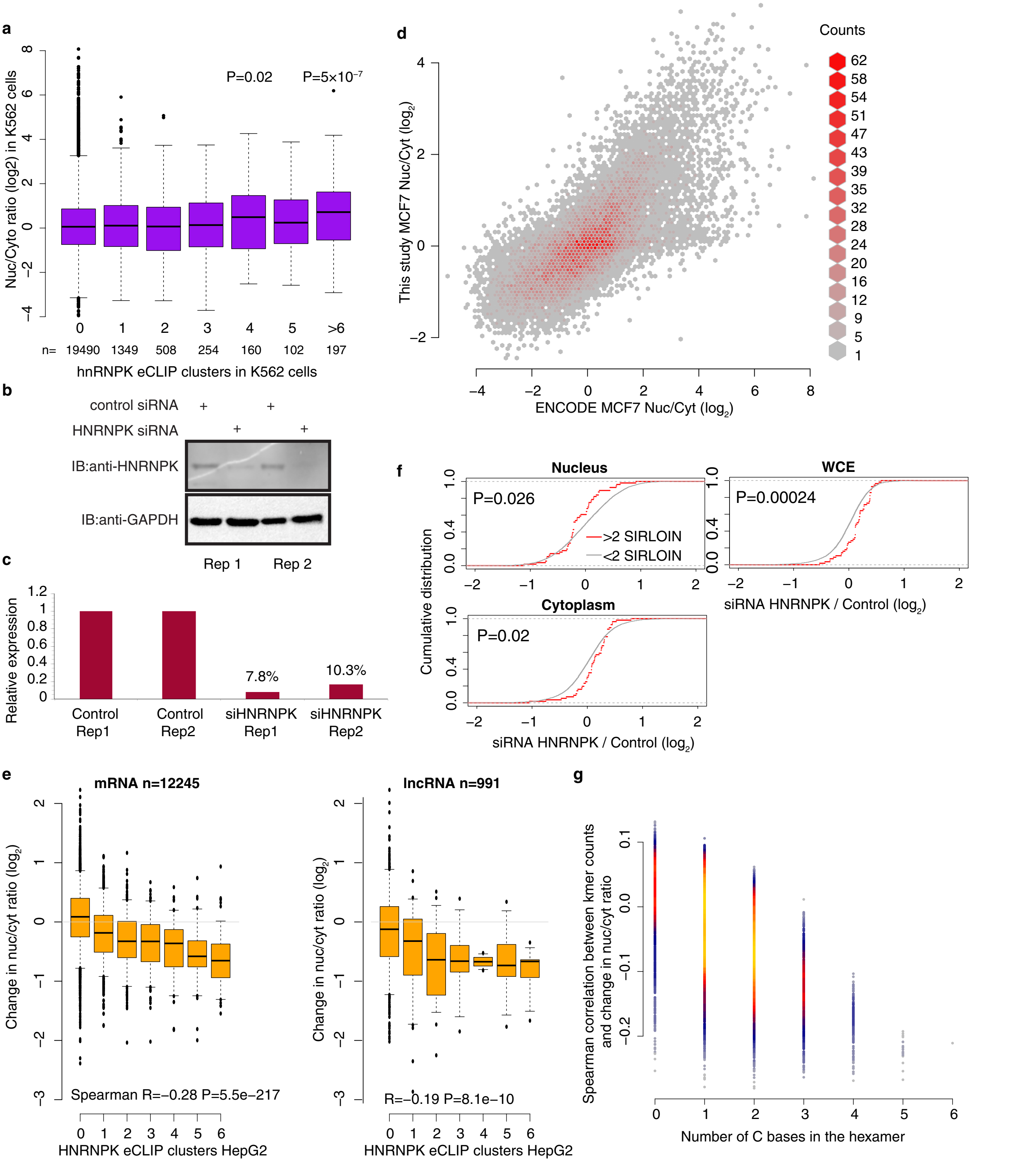
Extended Data Figure 3 | Comparison of splicing isoforms, chromatin enrichment and mRNA half-lives. a, Semi-quantitative RT–PCR in the indicated fraction (left) and gene structures (right) of genes with multiple alternative isoforms, some of which contain SIRLOIN elements. Gene structures taken from RefSeq, UCSC or GENCODE genes; Alu annotations are from the UCSC genome browser. Arrows indicate positions of the primers used for RT–PCR. RNA-seq coverage taken from MCF-7 RNA-seq from the ENCODE project. Experiments were repeated once. **b**, RNA-seq-based expression ratios comparing nuclear/cytoplasmic, chromatin/nucleoplasmic, and nucleolar/nucleoplasmic RNA fractions for RNAs with the indicated number of SIRLOIN or antisense SIRLOIN elements (data from the ENCODE project). *n* = 82–17,481 genes per group. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles. *P* values computed using two-sided Wilcoxon test. **c**, Comparison of half-lives in MCF-7 cells[9] of protein-coding genes with the indicated number of SIRLOIN elements. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles.



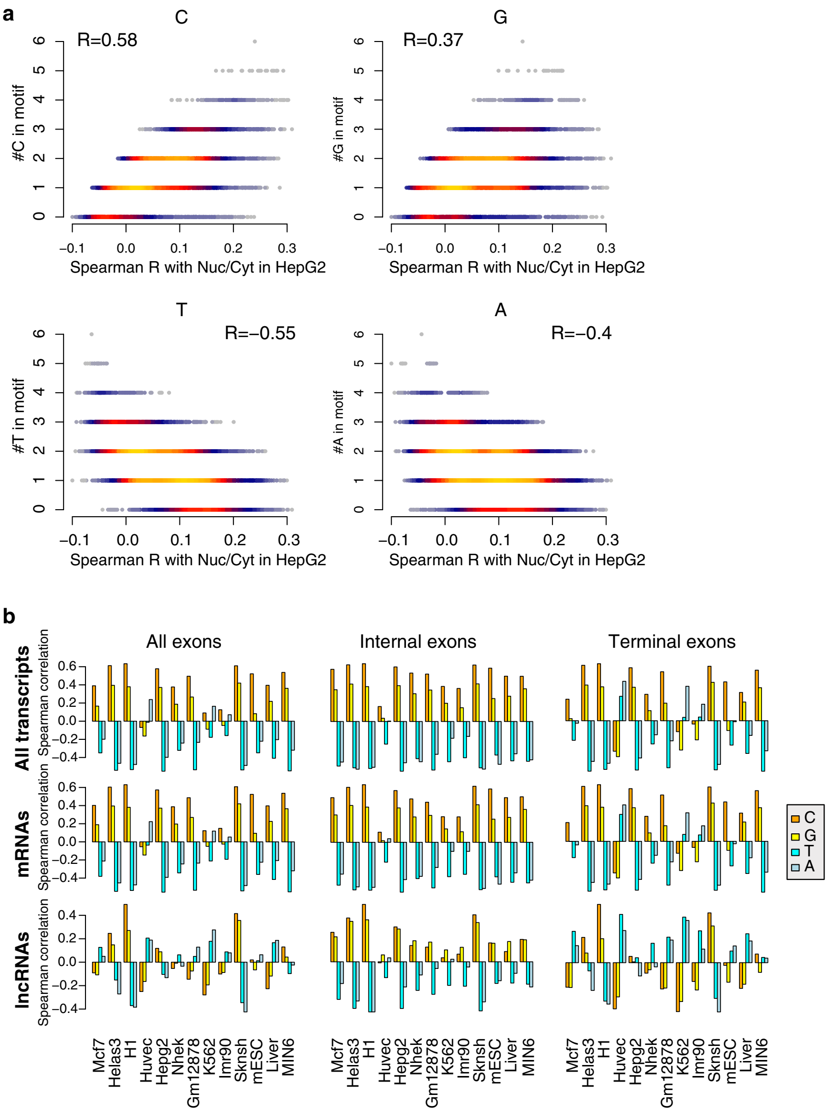
Extended Data Figure 4 | Nuclear/cytoplasmic ratios for NucLibB tiles. Effects of all the tiles in the indicated lncRNAs and mRNAs on nuclear/cytoplasmic expression ratios when cloned into the 3′ UTR of *AcGFP*. Tiles overlapping other repetitive elements are in black, and other tiles are in grey. Regions of consecutive tiles overlapping SIRLOIN elements are shaded grey.



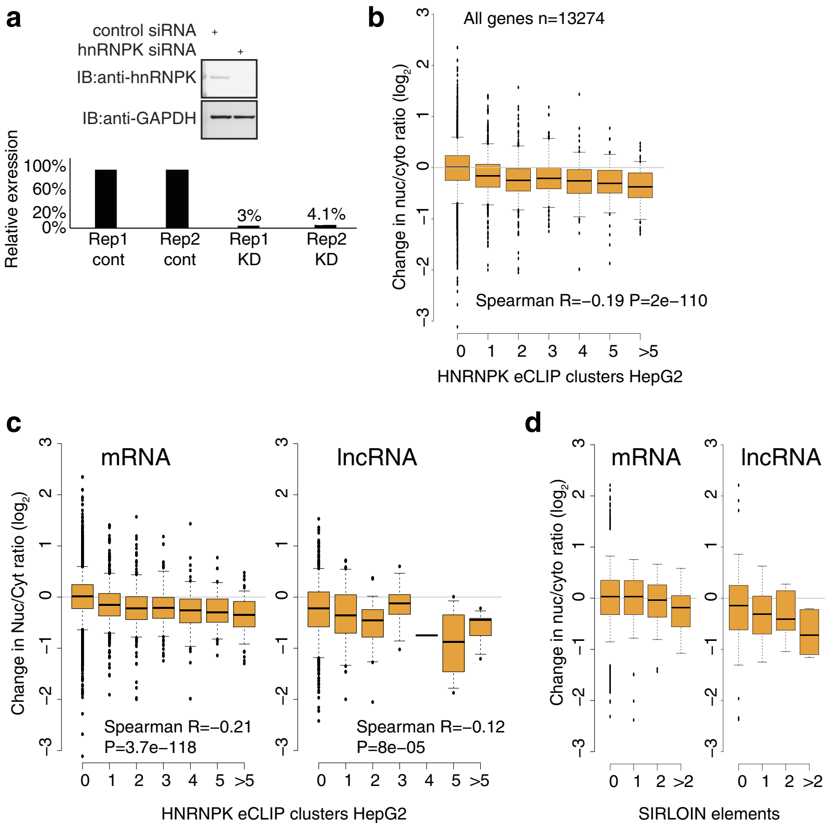
Extended Data Figure 5 | Statistics and mutagenesis results in NucLibB. **a**, Correlation between nuclear/cytoplasmic ratios and expression levels of sequences in NucLibB (only wild-type sequences were analysed and two replicates were pooled; $n = 2{,}930$ for all data and $n = 104$ for SIRLOIN data). SIRLOIN-containing sequences are circled. Colouring indicates the local point density. Both replicates are plotted together and correlations and P values were computed using Spearman correlation. **b**, **c**, Nuc/Cyt ratios and expression levels of 109-nt fragments containing repetitions of the indicated motifs from JPX#9 (**b**) and NucLibA PVT1#22 (**c**), separated by AT dinucleotides. Horizontal lines indicate levels of NucLibB JPX#9 (**b**) and PVT1#23 (**c**) sequence (the closest to the NucLibA PVT1#22 sequence). **d**, Effects of mutations in JPX#9 (2nd replicate) on localization (top) and expression levels (bottom). **e**, Effects of mutations in PVT1#22. Note that the wild-type sequence of PVT1#22 from NucLibA was not included in NucLibB, and so the values of PVT1#23, which is the closest to that sequence in NucLibB, are shown. **f**, Effect of shuffles of the indicated kmers in PVT1#22 sequence. **g**, Predicted secondary structures of full-length *AcGFP* mRNA with the indicated inserts. Secondary structure prediction by the Vienna package RNAfold server with default parameters. The SIRLOIN elements are indicated in blue.



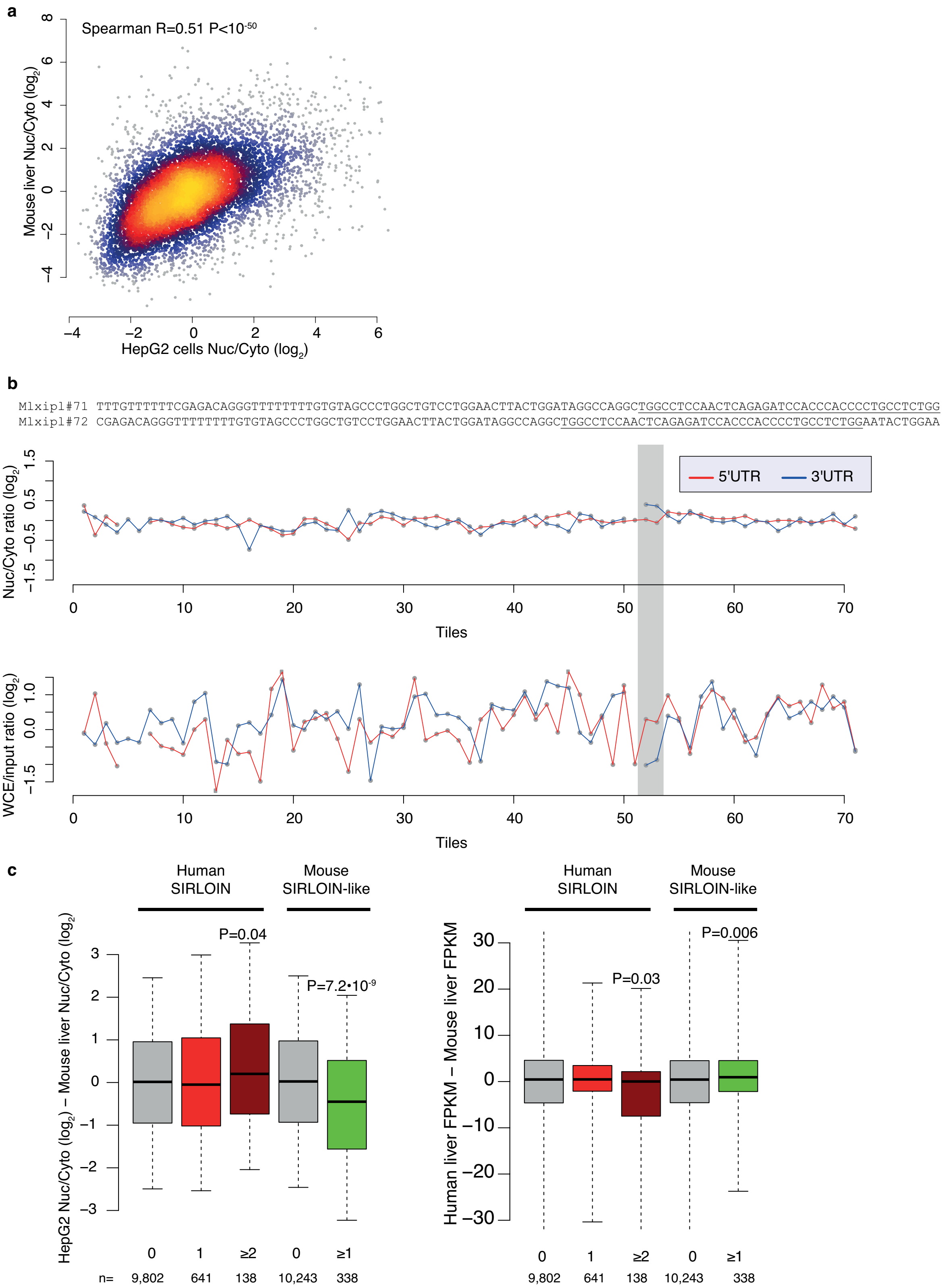
Extended Data Figure 6 | HNRNPK regulation in K562 and MCF7 cells. **a**, Ratios of expression levels in the nucleus/cytoplasm for transcripts with the indicated number of eCLIP clusters for HNRNPK in K652 cells. *P* values indicate significant difference by two-sided Wilcoxon test between genes with the indicated number of clusters and genes without eCLIP clusters. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles. **b**, Knockdown of HNRNPK in MCF7 cells using siRNAs assessed by western blot. Experiment was performed twice and both replicates are shown. **c**, HNRNPK mRNA levels measured by qRT–PCR. Each bar presents a single independent experiment. **d**, Correlation between nuclear/cytoplasmic ratios in MCF7 cells in the ENCODE data and in the RNA-seq data collected in this study. Hexagons are coloured according to the number of genes. $n = 13,235$. **e**, Ratios of expression levels in the nucleus/cytoplasm for transcripts with the indicated number of eCLIP clusters for HNRNPK in HepG2 cells, separately for lncRNAs and mRNAs. Otherwise as in Fig. 4f. **f**, Effects of HNRNPK knockdown on expression levels in the indicated sample, for lncRNAs and mRNAs containing the indicated number of SIRLOIN elements. Otherwise as in Fig. 4f. **g**, Spearman correlation coefficients between the number of appearances of hexamers in the internal exons of transcripts and the change in nuclear/cytoplasmic ratios following HNRNPK knockdown (average of two replicates). Hexamers are grouped by the number of C bases in the hexamer. Colouring indicates the local point density.



Extended Data Figure 7 | Correlations between hexamer occurrences and nuclear/cytoplasmic ratios in human and mouse cells. a, Spearman correlation coefficients between hexamers with the indicated number of bases and nuclear/cytoplasmic ratios in human HepG2 cells. Each point represents a hexamer sequence, the occurrences of which were counted in all exons of all >200-nt RefSeq transcripts. Colouring indicates the local point density. *R* values indicate Spearman correlation coefficients between the number of indicated bases in the hexamer and the nuclear/cytoplasmic ratios. **b**, Spearman correlation coefficients computed as in **a** for each of ten human ENCODE cell lines and three mouse cell types, when examining the indicated transcript types and the indicated exon subsets.

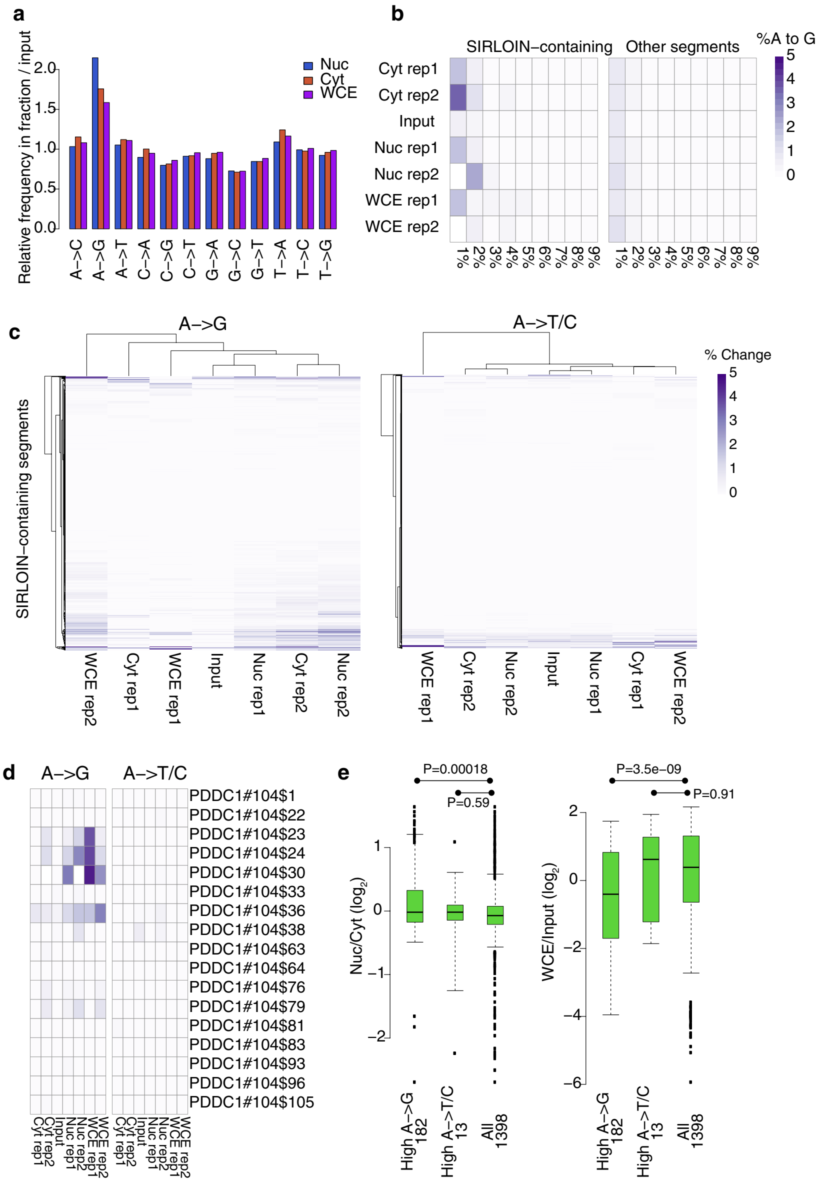


Extended Data Figure 8 | HNRNPK knockdown in HeLa cells. **a**, Representative western blot (top) and qRT–PCR quantification (bottom) of HNRNPK knockdown (each bar shows the level in a single experiment). Experiment was repeated twice. **b**, Changes in nuclear/cytoplasmic ratios of genes with the indicated number of HNRNPK eCLIP peaks following HNRNPK knockdown (DESeq2 analysis of two independent replicates). Otherwise as in Fig. 4f. *n* = 91–11,016 genes. **c**, As in **b**, separately for mRNA and lncRNAs. *n* = 85–10,077 mRNAs per group, 6–939 lncRNAs per group. **d**, Changes in nuclear/cytoplasmic ratios of mRNAs and lncRNAs with the indicated number of SIRLOIN elements. Otherwise as in Fig. 4f. *n* = 53–11,172 mRNAs per group, 10–891 lncRNAs per group.



```
Mlxipl#71 TTTGTTTTTTCGAGACAGGGTTTTTTTTGTGTAGCCCTGGCTGTCCTGGAACTTACTGGATAGGCCAGGCTGGCCTCCAACTCAGAGATCCACCCACCCCTGCCTCTGG
Mlxipl#72 CGAGACAGGGTTTTTTTTGTGTAGCCCTGGCTGTCCTGGAACTTACTGGATAGGCCAGGCTGGCCTCCAACTCAGAGATCCACCCACCCCTGCCTCTGGAATACTGGAA
```

Extended Data Figure 9 | SIRLOIN-like element in mouse. a, Comparison of nuclear/cytoplasmic ratios for orthologous genes in human HepG2 cells (ENCODE data) and mouse liver[3]. $n = 10{,}160$ conserved genes. Colouring indicates the local point density. Correlation computed using Spearman correlation. **b**, Top, sequences of the Mlxipl#71/72 tiles in NucLibA. Centre, nuclear/cytoplasmic ratios for each tile with a sufficient number of reads in NucLibA. Bottom, WCE/input expression levels for each tile. The region of tiles #71–72 is shaded in grey. **c**, Left, difference in log-transformed nuclear/cytoplasmic ratios between orthologous genes with the indicated number of human SIRLOIN elements or mouse SIRLOIN-like elements. Right, differences in overall expression levels; human liver expression data taken from the Human Proteome Atlas[39], mouse liver expression data taken from the ENCODE project. *P* values computed using two-sided Wilcoxon test, comparing to genes without SIRLOIN elements. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles.



Extended Data Figure 10 | RNA editing in SIRLOIN elements. **a**, Frequencies of all base substitutions in each fraction, normalized to their occurrence in the input libraries. Samples from NucLibB were analysed, and all these samples were sequenced in the same NextSeq sequencing run. The number of substitutions was tallied for reads with fewer than six mutations relative to the reference library sequence. The count of each mutation type in each sample was then divided by the total number of mutations in the sample, the two replicates were averaged and the number of mutations in each sample type was divided by the number in the input library. **b**, For each reference A base in the indicated sample and the indicated group of segments, we computed the fraction of reads that contained a mutation to a G. The positions were then binned in 1% bins, and the heat map shows that fraction of bases mapping to each bin. In all cases, >90% of As were found in the <1% editing bin, which is not shown. Each row corresponds to a single experiment. **c**, In each heat map, each row corresponds to an A in a segment containing a SIRLOIN element, and the fraction of reads containing an A→G or A→T/C mutation is shown in each sample. Each column corresponds to a single experiment. **d**, As in **c**, but showing a specific SIRLOIN-containing segment. **e**, Correlation between high levels of RNA editing, localization (left) and expression levels (right). ‘High A→X’ are those segments in which the mean A→X mutation levels were at least three times higher than the median across all segments, in at least three of the four libraries (nucleus/cytoplasm, two replicates). Numbers near each group name correspond to sample size. Nuc/Cyt and WCE/Input levels were averaged across the two replicates. Boxplots show 5th, 25th, 50th, 75th and 95th percentiles and *P* values were computed using a two-sided Wilcoxon test.

LETTER

doi:10.1038/nature25507

Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress

Morgane Macheret[1] & Thanos D. Halazonetis[1]

Oncogene-induced DNA replication stress contributes critically to the genomic instability that is present in cancer[1–4]. However, elucidating how oncogenes deregulate DNA replication has been impeded by difficulty in mapping replication initiation sites on the human genome. Here, using a sensitive assay to monitor nascent DNA synthesis in early S phase, we identified thousands of replication initiation sites in cells before and after induction of the oncogenes *CCNE1* and *MYC*. Remarkably, both oncogenes induced firing of a novel set of DNA replication origins that mapped within highly transcribed genes. These ectopic origins were normally suppressed by transcription during G1, but precocious entry into S phase, before all genic regions had been transcribed, allowed firing of origins within genes in cells with activated oncogenes. Forks from oncogene-induced origins were prone to collapse, as a result of conflicts between replication and transcription, and were associated with DNA double-stranded break formation and chromosomal rearrangement breakpoints both in our experimental system and in a large cohort of human cancers. Thus, firing of intragenic origins caused by premature S phase entry represents a mechanism of oncogene-induced DNA replication stress that is relevant for genomic instability in human cancer.

Studies addressing the mechanisms that underlie oncogene-induced DNA replication stress have implicated several possible causes: reduced or increased origin firing, depletion of nucleotides,

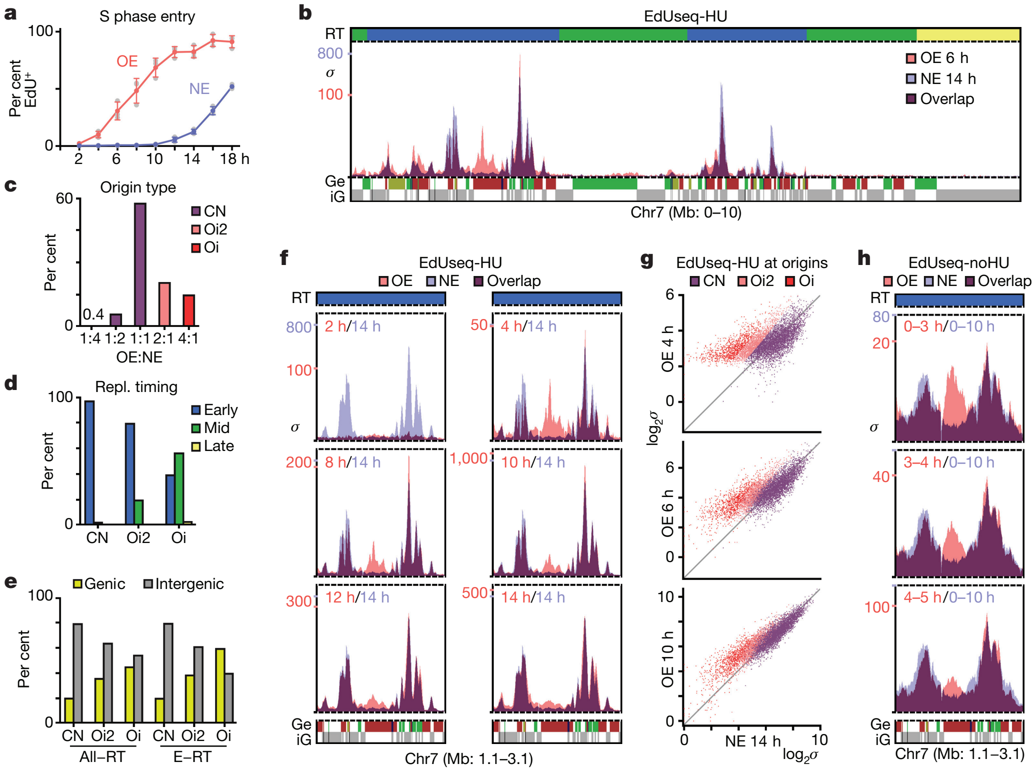


Figure 1 | Firing of novel origins upon cyclin E overexpression. **a**, Percentage of EdU-positive cells (mean and s.d.; $n = 3$ independent experiments; grey symbols, individual data points) at the indicated times after mitotic shake-off. OE, overexpression of cyclin E; NE, normal levels of cyclin E. **b**, Replication initiation profiles (EdUseq-HU) at a representative genomic region in OE and NE cells collected 6 and 14 h after mitotic shake-off, respectively. RT, replication timing (blue, early; green, mid; yellow, late S phase); Ge, genes (green, forward direction of transcription; red, reverse; yellow, unspecified; blue, multiple genes within bin); iG, intergenic (grey). Bin resolution, 10 kb; ruler scale, 100 kb; σ, sigma (normalized number of sequence reads per bin divided by its s.d.). **c**, Classification of origins based on adjusted σ value ratios in OE over NE cells: CN, constitutive (less than twofold induction); Oi2, oncogene-induced 2 (more than twofold induction); Oi, oncogene-induced (more than fourfold induction). **d**, **e**, Distribution of CN, Oi2 and Oi origins according to RT (**d**) and gene annotation (**e**) (All-RT, all RT domains; E-RT, early S phase RT genomic domains). **f**, **g**, Replication initiation profiles (EdUseq-HU) at a representative genomic region (**f**) and scatter plots of EdUseq-HU values at all origins (**g**) in NE and OE cells at the indicated times after mitotic shake-off. **h**, Replication initiation profiles (EdUseq-noHU) at a representative genomic region in OE and NE cells. EdU was present during the indicated times following mitotic shake-off.

[1]Department of Molecular Biology, University of Geneva, 1205 Geneva, Switzerland.

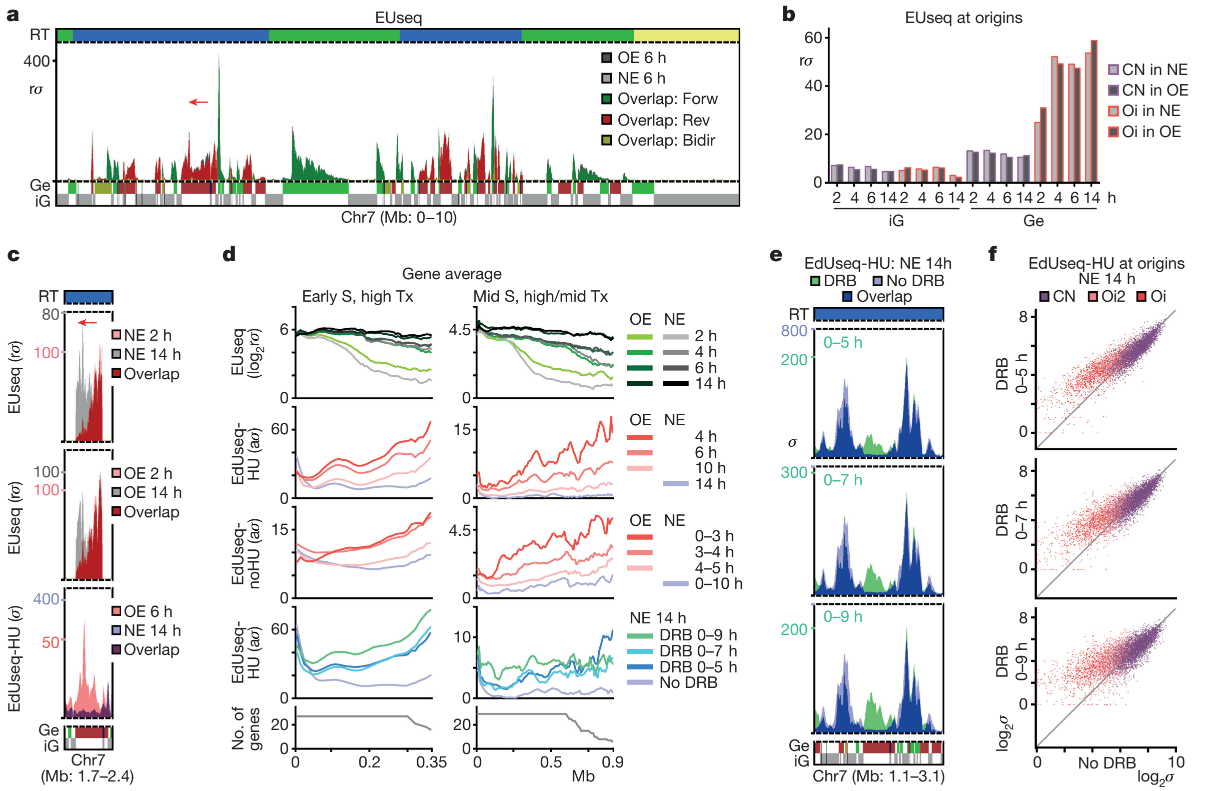


Figure 2 | Suppression of intragenic origin firing by transcription. **a**, Newly synthesized transcript profiles (EUseq) at a representative genomic region in OE and NE cells 6 h after mitotic shake-off (OE: dark grey; NE: light grey; overlap: colour; direction of transcripts: green, forward (forw); red, reverse (rev); yellow, bidirectional (bidir)). Replication timing (RT) and gene (Ge/iG) annotations are as in Fig. 1b. rσ, relative sigma. **b**, Median transcript levels (EUseq) at the genomic bins corresponding to genic (Ge) and intergenic (iG), constitutive (CN) and oncogene-induced (Oi) origins in NE (light grey) and OE (dark grey) cells at the indicated times after mitotic shake-off. **c**, Transcription (EUseq) and replication initiation (EdUseq-HU) profiles at the indicated times after mitotic shake-off for the gene marked by the arrow in **a**. **d**, Average transcription (EUseq) and replication initiation (EdUseq-HU or EdUseq-noHU) along the length of large (>0.35 Mb for early S; >0.65 Mb for mid S), transcribed (Tx) genes in OE, NE and DRB-treated NE cells at various time points after mitotic shake-off. High Tx, upper tercile; mid Tx, middle tercile; aσ, adjusted sigma; no. of genes, number of averaged genes. **e**, **f**, Replication initiation (EdUseq-HU) profiles at a representative genomic region (**e**) and scatter plots of EdUseq-HU values at all origins (**f**) in control (no DRB) and DRB-treated (0–5, 0–7 or 0–9 h) NE cells collected 14 h after mitotic shake-off.

shortage of replication factors, reduced fork elongation rates, increased transcription and replication–transcription conflict[5–13]. However, to our knowledge, no study has yet mapped genome-wide DNA replication and transcription in cells, before and after oncogene activation, which could provide important insights into oncogene-induced DNA replication stress.

We studied U2OS human cell lines, in which inducible activation of either of the proto-oncogenes *CCNE1* (cyclin E) or *MYC* leads to DNA replication stress[6,13–15]. Amplifications of the *CCNE1* and *MYC* genes are among the most frequent genetic changes in human cancers[2–4]. We first focused on the cyclin E system. As previously shown[5,14,16], overexpression of cyclin E shortened the length of the G1 cell cycle phase, from about 10–12 h for cells with normal cyclin E activity, to as little as 2–4 h (Fig. 1a and Extended Data Fig. 1). To examine DNA replication initiation (origin firing), cells with normal levels of cyclin E and cells overexpressing cyclin E were collected by mitotic shake-off and allowed to proceed through the cell cycle in the presence of 5-ethynyl-2′-deoxyuridine (EdU) and hydroxyurea (HU)[17–19]. EdU-labelled DNA was isolated and subjected to high-throughput sequencing (EdUseq-HU; Extended Data Fig. 1b, c). The data, analysed at a genomic bin resolution of 10 kb, yielded well-resolved peaks, corresponding to the regions where DNA replication initiated (Fig. 1b and Supplementary Table 1). Of 6,164 identified peaks, 927 were induced strongly (at least fourfold) in cyclin E-overexpressing cells and will be referred to as oncogene-induced (Oi) origins; 1,281 were induced modestly by cyclin E overexpression and will be referred to as Oi2 origins; and 3,956 were of similar magnitude in the normal and cyclin E-overexpressing cells and will be referred to as constitutive origins (Fig. 1b, c).

To determine whether constitutive and Oi origins were qualitatively different, we examined their genomic distribution in relation to replication timing and gene annotation. Early, mid and late S phase replicating domains were mapped by REPLIseq[20] (Extended Data Fig. 2). The constitutive origins were present exclusively in early S domains, as expected of cells that had just entered S phase, whereas the Oi origins exhibited a broader distribution that encompassed both early and mid S domains (Fig. 1d). In regard to gene annotation, while the constitutive origins mapped predominantly to intergenic regions, a substantial fraction of the Oi origins, particularly those in early S domains, mapped within protein-coding genes (Fig. 1e). Similar results were obtained when the resolution of origin mapping was increased from 10 kb to 1 kb by treating the cells with mimosine or aphidicolin, in addition to hydroxyurea, to more robustly arrest fork progression after origin firing (Extended Data Fig. 3 and Supplementary Table 2).

We examined origin firing at multiple time points after mitotic exit in cyclin E-overexpressing cells, and found that Oi origins fired predominantly in the cells with the shortest G1 phases (less than 6 h; Fig. 1f, g and Supplementary Table 3). As expected, origin firing was not observed before S phase entry (that is, within the first 2 h after mitotic exit). Furthermore, the Oi origins initiated DNA replication even in the absence of hydroxyurea, indicating that they were not dormant origins (Fig. 1h and Extended Data Fig. 3f). Thus, oncogene activation

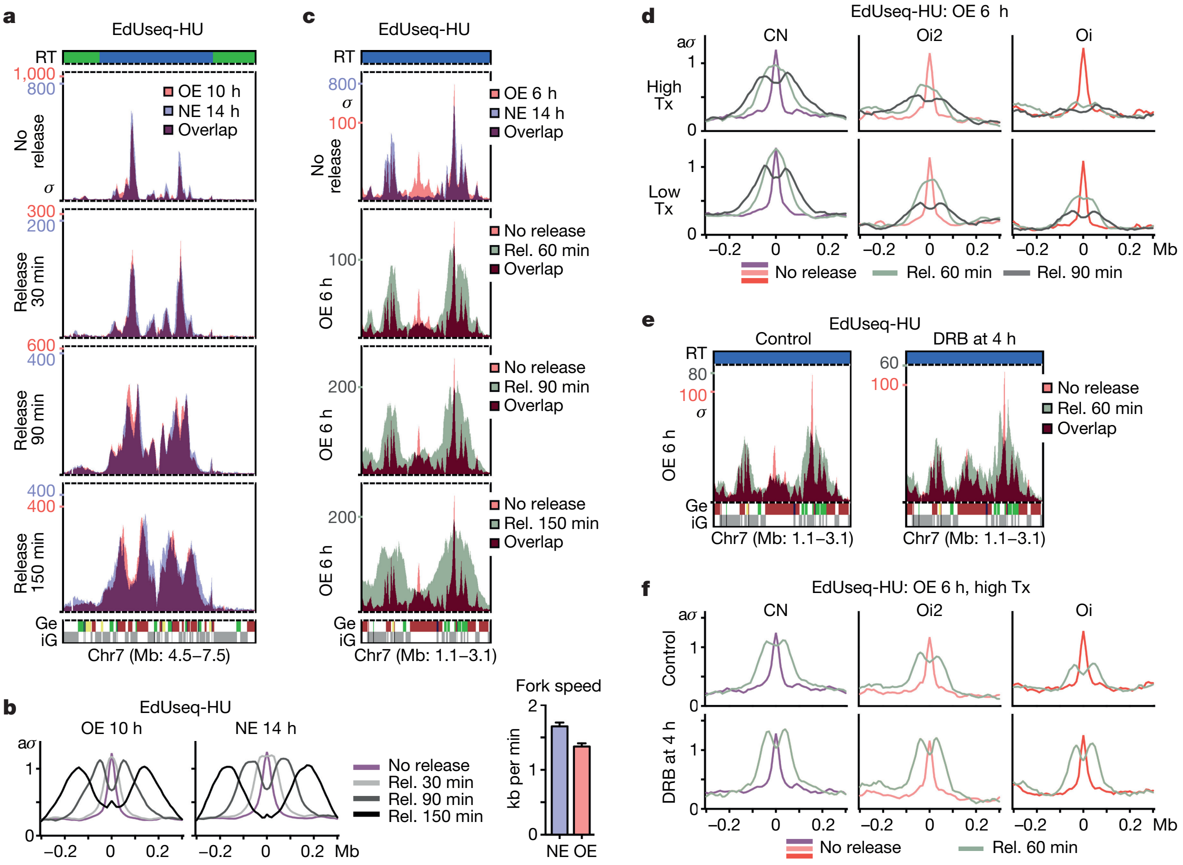


Figure 3 | Collapse of Oi forks due to conflicts with transcription. **a**, Fork progression profiles (EdUseq-HU/release) at a representative genomic region in OE and NE cells that were arrested with hydroxyurea for 10 or 14 h, respectively, after mitotic shake-off and then released for the indicated time periods. EdU was added 30 min before the cells were collected. The no release (EdUseq-HU) profile is shown as a reference. RT and gene (Ge/iG) annotations are as in Fig. 1b. **b**, Genome-wide average fork progression at constitutive origins in OE and NE cells from the experiment shown in **a**. Fork speeds were determined from the distances travelled by the forks between the 90 and 150 min release (rel) time points. **c**, Replication initiation (EdUseq-HU) profiles at a representative genomic region in OE and NE cells (top panel, reference) and fork progression profiles of OE cells released from a 6 h hydroxyurea block for the indicated times and labelled with EdU for 30 min before collection (lower panels, EdUseq-HU/release). **d**, Genome-wide average fork progression at constitutive and oncogene-induced origins located in high- or low-transcribed (Tx) regions (upper and lower terciles, respectively) from the experiment shown in **c**. **e**, Rescue of fork progression (EdUseq-HU/release) at Oi origins within a representative genomic region in DRB-treated OE cells. DRB was added 4 h after mitotic shake-off and kept until collection. Control, no DRB. **f**, Genome-wide average fork progression at CN, Oi2 and Oi origins located in highly transcribed regions from the experiment shown in **e**.

led to the firing of novel replication origins within genomic domains normally devoid of replication initiation in cells that entered S phase prematurely.

The aberrant firing of intragenic Oi origins could be related to deregulation of transcription, but the newly synthesized transcript profiles (EUseq) of cells examined at multiple time points after mitotic exit showed that cyclin E overexpression did not affect transcription genome-wide (Fig. 2a and Extended Data Fig. 4a) or at the genomic sites where the Oi origins fired (Fig. 2b and Extended Data Fig. 3e). Notably, whereas the constitutive origins, including the genic ones, mapped to non-transcribed or weakly transcribed regions, the genic Oi origins mapped to sites that were highly transcribed at all time points examined, except for the early time point of 2 h after mitotic exit (Fig. 2b and Extended Data Fig. 3e). Thus, replication at the sites of Oi origins may initiate before these sites are transcribed. Indeed, specific examples (Fig. 2c) and averages of transcription and replication along large genes (Fig. 2d and Extended Data Fig. 4b) revealed that, 2 h after mitotic exit, the transcription wave front had not yet reached the 3′ ends of genes, where most of the Oi origins were located.

Transcription has been proposed to inactivate intragenic origins before S phase entry[21,22]. Thus, one interpretation of our results is that intragenic origins fired upon oncogene-induced premature S phase entry because transcription did not have the time needed to reach the end of the transcription units. To test this hypothesis, we used 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) to inhibit transcription elongation for the first 5, 7 or 9 h of G1 in cells with normal levels of cyclin E. The DRB treatment did not decrease the length of G1, but it reduced the time during which transcription was active in G1 (Extended Data Fig. 4c–f), and, consistent with our hypothesis, it led to firing of intragenic origins at the same genomic positions as the Oi origins (Fig. 2d–f).

Inducible activation of the proto-oncogene *MYC* in U2OS cells led to a similar shortening of the G1 phase and firing of intragenic Oi origins, many of which overlapped with the cyclin E-induced Oi origins (Extended Data Fig. 5). Myc activation affected transcription more broadly than cyclin E overexpression, but, as observed in the cyclin E system, constitutive and Oi origins were associated with sites of low and high transcriptional activity, respectively, highlighting the mechanistic similarities between cyclin E and Myc-induced intragenic origin firing (Extended Data Fig. 5i–k).

HeLa cells, a well-characterized cancer cell line, entered S phase with similar kinetics to those of U2OS cells overexpressing cyclin E. Analysis of DNA replication initiation at 6 and 14 h after mitotic shake-off revealed firing of the same Oi origins as characterized in the U2OS cell systems, for the cells with the shortest G1 phases (Extended Data

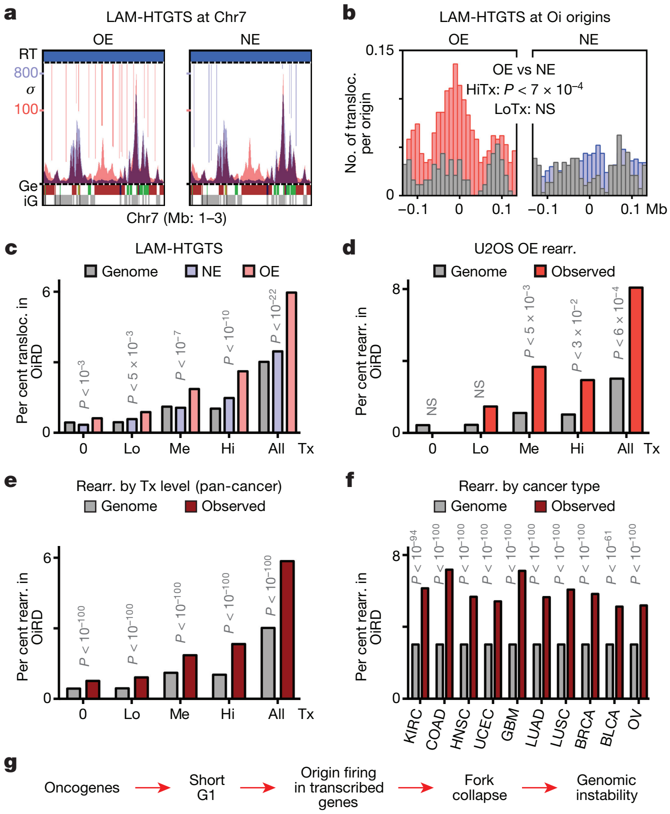


Figure 4 | Oi origins are associated with DNA DSB formation and genomic rearrangements. a, Translocations identified by LAM-HTGTS within a representative genomic region in OE and NE cells shown as vertical lines (red and blue, respectively, with colour intensity reflecting the number of translocations per genomic bin) in the context of the replication initiation profiles of these cells (EdUseq-HU). **b**, Number of translocations per origin in OE and NE cells at oncogene-induced origins and surrounding genomic bins. Data are plotted separately for origins mapping to high- (red, blue; HiTx) and low-transcribed (grey; LoTx) sites (upper and lower terciles, respectively) and statistical comparisons between OE and NE samples were performed by random permutation tests. **c**–**f**, Mapping of translocations (Transloc; $n = 27{,}364$) identified by LAM-HTGTS (**c**), genomic rearrangement (Rearr; $n = 136$) breakpoints identified previously[14] in the same OE cells (**d**) and genomic rearrangement (Rearr; $n = 490{,}711$) breakpoints from a TCGA pan-cancer data set[4] (**e**, **f**) to genomic regions replicated from Oi origins (oncogene-induced replication initiation domains, OiRDs). The fraction of translocations or breakpoints mapping to OiRDs is shown for non-transcribed (0), low- (Lo), medium- (Me) and high-transcribed (Hi) genomic bins, and for all genomic bins (**c**–**e**) or for all genomic bins in common cancer types (**f**). The distribution of OiRDs in the genome is shown in grey. Statistical comparisons are between NE (blue) and OE (pink) samples (**c**) or between observed (red) and genomic (grey) frequencies (**d**–**f**) and were performed either by random permutation tests (**c**, **d**) or by calculating z-scores (**e**, **f**). NS, not significant. KIRC, kidney renal cell; COAD, colon adenocarcinoma; HNSC, head and neck squamous cell; UCEC, uterine cervix; GBM, glioblastoma multiformae; LUAD, lung adenocarcinoma; LUSC, lung squamous cell; BRCA, breast; BLCA, bladder; OV, ovary. **g**, Proposed mechanism for oncogene-induced DNA replication stress.

Figs 6, 7). By contrast, Oi origins did not fire in RPE1 cells, which are not transformed and which have long G1 phases (Extended Data Figs 6, 7). Notwithstanding the differences in Oi origin firing, the majority of constitutive origins were shared in all cell lines examined (Extended Data Figs 5h, 6, 7).

To determine whether Oi origin firing could lead to DNA replication stress, we compared fork progression from constitutive and Oi origins. To monitor constitutive origins, cells expressing normal or high levels of cyclin E were arrested with hydroxyurea for 14 and 10 h after mitotic exit, respectively, and then released from the hydroxyurea block for various time periods (30 min–2.5 h) with EdU being added for the last 30 min before harvesting the cells (EdUseq-HU/release) (Fig. 3a). There was robust incorporation of EdU label around the constitutive origins at all release time points examined. Although fork progression from constitutive origins was slower in cyclin E-overexpressing cells, which has been previously attributed to the increased number of origins[6,7], there was no evidence of fork collapse (Fig. 3a, b). To monitor Oi origins, we repeated the experiment with cyclin E-overexpressing cells that were arrested with hydroxyurea for only 6 h after mitotic exit, a time point at which Oi origins have fired with high efficiency. Notably, the Oi origins failed to recover from the hydroxyurea block, indicating fork collapse, whereas in the same cells forks from constitutive origins did not collapse (Fig. 3c, d and Supplementary Table 4). The degree of fork collapse correlated with the level of transcription at the sites of Oi origins (Fig. 3d), and treating the cells with the transcription inhibitor DRB as the cells were entering S phase rescued fork collapse (Fig. 3e, f and Supplementary Table 4), suggesting that replication–transcription conflicts were the underlying cause. This conclusion was further supported by analysis of cells with normal levels of cyclin E, in which Oi origins were induced to fire by inhibiting transcription in early G1 (Extended Data Fig. 8).

We next examined whether fork collapse at Oi origins led to the formation of DNA double-stranded breaks (DSBs). We studied unsynchronized cells expressing normal or high levels of cyclin E in the absence of exogenous replication stress-inducing agents, such as hydroxyurea. DNA DSBs leading to translocations with a CRISPR–Cas9-induced site-specific DNA DSB were identified using the linear amplification-mediated high-throughput genome-wide translocation sequencing (LAM-HTGTS) assay[23]. Translocations were mapped to Oi origins and also, more broadly, to the genomic domains replicated from these origins (Oi replication initiation domains), which we identified from the replication initiation profiles of cells that were not treated with hydroxyurea (Fig. 1h). Translocation breakpoints were enriched at Oi origins specifically in cyclin E-overexpressing cells (Fig. 4a, b). The effect was dependent on the level of transcription, as only Oi origins at highly transcribed loci were significantly associated with translocations (Fig. 4b). Similarly, the percentage of translocations mapping to Oi replication initiation domains increased significantly upon cyclin E overexpression, and the identified translocations mapped preferentially to those domains with mid or high levels of transcription (Fig. 4c and Extended Data Fig. 9a). Similar findings were obtained by examining breakpoints of previously identified gross chromosomal rearrangements (amplifications and deletions) in the same cyclin E-inducible cells[14] and in a large cohort of human cancers[4] (Fig. 4d–f, Extended Data Fig. 9b–d and Supplementary Table 5).

The genome-wide replication initiation and transcription profiles described here provide new insights into how oncogenes induce DNA replication stress. We observed that oncogenes induce the firing of novel replication origins, which, unlike the constitutive origins, are intragenic and give rise to replication forks that are prone to collapse. Therefore, oncogene-induced DNA replication stress does not involve all replisomes, but the subset derived from the novel, intragenic origins. This latter subset was also associated with a higher frequency of genomic rearrangements in cancer. Our study did not interrogate the late-replicating part of the genome, which is where most common fragile sites map[24–26]. Nevertheless, more chromosomal rearrangements in human cancers map to early- than to late-replicating genomic domains (Extended Data Fig. 9e), and early replicating fragile sites have also been identified[27]. The collapse of forks initiating from intragenic, oncogene-induced origins could be attributed to replication–transcription conflicts, whereas forks from intergenic, constitutive origins did not collapse, even when replicating highly transcribed genes. This difference in behaviour may relate to the fact that head-on replication–transcription collisions, the most damaging type[28], cannot be prevented when origins fire within genes. However, for constitutive origins, which are intergenic, head-on collisions can be avoided by a genomic organization, including replication fork barriers, that favours co-directionality of replication with transcription[29,30]. Alternatively, forks might be particularly sensitive to replication–transcription conflicts shortly after origin firing (for example, before lagging strand synthesis has converted the initial origin bubble to double-stranded DNA daughter molecules), which would explain why forks from intragenic origins are prone to collapse.

Our study also helps to explain how oncogenes induce firing of intragenic origins. We observed that transcription suppresses origin firing from within genes. In normal cell cycles, the length of G1 is sufficient for transcription to inactivate origins across the entire length of genes. However, oncogenes greatly reduce the length of G1 and therefore leave insufficient time for transcription to inactivate all intragenic origins (Fig. 4g and Extended Data Fig. 10). This concept of transcription erasing intragenic origins fits with the well-known observation that the largest genes in the human genome are late replicating, which would, in principle, provide more time for transcription to reach the 3′ end of these genes before they initiate replication. This mechanism also helps to reconcile how shortening of the G1 phase, a typical outcome of oncogene activation, leads to aberrant origin firing and DNA replication stress.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 February 2017; accepted 9 January 2018.
Published online 21 February 2018.

1. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319,** 1352–1355 (2008).
2. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
3. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
4. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
5. Ekholm-Reed, S. *et al.* Deregulation of cyclin E in human cells interferes with prereplication complex assembly. *J. Cell Biol.* **165,** 789–800 (2004).
6. Jones, R. M. *et al.* Increased replication initiation and conflicts with transcription underlie cyclin E-induced replication stress. *Oncogene* **32,** 3744–3753 (2013).
7. Beck, H. *et al.* Cyclin-dependent kinase suppression by WEE1 kinase protects the genome through control of replication initiation and nucleotide consumption. *Mol. Cell. Biol.* **32,** 4226–4236 (2012).
8. Di Micco, R. *et al.* Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* **444,** 638–642 (2006).
9. Bester, A. C. *et al.* Nucleotide deficiency promotes genomic instability in early stages of cancer development. *Cell* **145,** 435–446 (2011).
10. Aird, K. M. *et al.* Suppression of nucleotide metabolism underlies the establishment and maintenance of oncogene-induced senescence. *Cell Reports* **3,** 1252–1265 (2013).
11. Toledo, L. I. *et al.* ATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell* **155,** 1088–1103 (2013).
12. Kotsantis, P. *et al.* Increased global transcription activity as a mechanism of replication stress in cancer. *Nat. Commun*. **7,** 13087 (2016).
13. Bartkova, J. *et al.* Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature* **444,** 633–637 (2006).
14. Costantino, L. *et al.* Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343,** 88–91 (2014).
15. Maya-Mendoza, A. *et al.* Myc and Ras oncogenes engage different energy metabolism programs and evoke distinct patterns of oxidative and DNA replication stress. *Mol. Oncol*. **9,** 601–616 (2015).
16. Resnitzky, D., Gossen, M., Bujard, H. & Reed, S. I. Acceleration of the G1/S phase transition by expression of cyclins D1 and E with an inducible system. *Mol. Cell. Biol.* **14,** 1669–1679 (1994).
17. Katou, Y. *et al.* S-phase checkpoint proteins Tof1 and Mrc1 form a stable replication-pausing complex. *Nature* **424,** 1078–1083 (2003).
18. MacAlpine, D. M., Rodríguez, H. K. & Bell, S. P. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev.* **18,** 3094–3105 (2004).
19. Karnani, N. & Dutta, A. The effect of the intra-S-phase checkpoint on origins of replication in human cells. *Genes Dev*. **25,** 621–633 (2011).
20. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107,** 139–144 (2010).
21. Sasaki, T. *et al.* The Chinese hamster dihydrofolate reductase replication origin decision point follows activation of transcription and suppresses initiation of replication within transcription units. *Mol. Cell. Biol*. **26,** 1051–1062 (2006).
22. Powell, S. K. *et al.* Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J.* **34,** 531–543 (2015).
23. Hu, J. *et al.* Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat. Protocols* **11,** 853–871 (2016).
24. Wilson, T. E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25,** 189–200 (2015).
25. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44,** 966–977 (2011).
26. Letessier, A. *et al.* Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470,** 120–123 (2011).
27. Barlow, J. H. *et al.* Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152,** 620–632 (2013).
28. Prado, F. & Aguilera, A. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *EMBO J.* **24,** 1267–1276 (2005).
29. Petryk, N. *et al.* Replication landscape of the human genome. *Nat. Commun.* **7,** 10208 (2016).
30. Martin, M. M. *et al.* Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome Res.* **21,** 1822–1832 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank U. Schibler, R. Pillai, M. Docquier and present and past laboratory members for helpful discussions; J. Bartek for the U2OS cells inducibly overexpressing cyclin E; M. Eilers for the U2OS MycER cells; R. Beroukhim and S. Schumacher for access to and help with TCGA cancer data sets; and N. Roggli for help with the graphics scripts and the Flow Cytometry and Genomics platforms of the University of Geneva. This work was supported by grants from the European Commission (ONIDDAC) and the Swiss Science National Foundation.

Author Contributions T.D.H. and M.M. designed the experiments and wrote the paper; M.M. performed the experiments; T.D.H. wrote the computer scripts and performed the bioinformatic analyses with contributions from M.M.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to T.D.H. (thanos.halazonetis@unige.ch).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture. U2OS cells inducibly overexpressing cyclin E (U2OS-CE; from J. Bartek), were maintained in Dulbecco's modified Eagle's medium (Invitrogen, Cat. No. 11960), supplemented with 10% fetal bovine serum (FBS; Invitrogen, Cat. No. 10500), penicillin/streptomycin/glutamine (Invitrogen, Cat. No. 10378-016), G418 400 μg/ml (Invitrogen, Cat. No. 10131-027), puromycin 1 μg/ml (Sigma, Cat. No. P8833) and tetracycline 2 μg/ml (Sigma, Cat. No. T7660). At the indicated number of days before the experiment, the cells were split into two aliquots. One aliquot was cultured in medium without tetracycline to induce cyclin E overexpression (OE cells) and the other in medium with 1 μg/ml doxycycline (Sigma, Cat. No. D3447) to maintain low levels of ectopic cyclin E expression (NE cells). U2OS cells inducibly activating Myc (U2OS-MycER; from M. Eilers) were maintained in Dulbecco's modified Eagle's medium without phenol red (Invitrogen, Cat. No. 31053), supplemented with 10% FBS and penicillin/streptomycin/glutamine. At the indicated number of days before the experiment, MycER activity was induced with 100 nM 4-hydroxytamoxifen (4-OHT; Sigma, Cat. No. H7904) dissolved in methanol. HeLa cells were purchased from the American Type Culture Collection (ATCC) and were maintained in Dulbecco's modified Eagle's medium, supplemented with 10% FBS and penicillin/streptomycin/glutamine. hTERT-RPE1 retinal pigment epithelial cells were purchased from the ATCC and were cultured in Dulbecco's Modified Eagle's Medium/Ham's F-12 (Invitrogen, Cat. No. 12634-010), supplemented with 10% FBS, penicillin/streptomycin/glutamine and hygromycin B (Sigma, Cat. No. H3274). All cell lines were identified by karyotyping and high-throughput genome sequencing and were routinely tested for the absence of mycoplasma contamination.

Antibodies, fluorescence and immunoblotting. Antibodies specific for cyclin E (Novocastra, Cat. No. NCL- CYCLINE), α-actinin (Millipore, Cat. No. 05-384) and c-Myc (Cell Signalling, Cat. No. 5605) were obtained from the indicated vendors. Immunofluorescence and immunoblotting were performed as previously described[14]. EU staining was performed using the Click-iT RNA Alexa Fluor 488 Imaging Kit (ThermoFisher, Cat. No. C10329) according to the manufacturer's instructions. Cell nuclei were counterstained with DAPI.

REPLIseq. U2OS-CE cells were cultured for 1, 2 or 7 days with or without doxycycline. The day before the experiment, the cells were re-seeded in order to obtain 70% confluency the following day. EdU (Invitrogen, Cat. No. A10044) was added at a concentration of 25 μM for 30 min to the asynchronously growing cells. The cells were then collected, fixed in 90% methanol overnight and permeabilized with 0.2% triton-X in PBS. EdU was coupled to a cleavable biotin linker (Azide-PEG(3+3)-S-S-biotin) (Jena Biosciences, Cat. No. CLK-A2112-10) using the Click-it Kit (Invitrogen Cat. No. C-10420). Genomic DNA was stained with propidium iodide (Sigma, Cat. No. 81845) in combination with RNase (Roche, Cat. No. 11119915001) and the cells were then sorted into three cell cycle populations according to DNA content using a MoFlo Astrios flow sorter (Beckman Coulter) at the Flow Cytometry platform of the Medical Faculty of the University of Geneva. DNA isolated from the sorted cells was purified by phenol-chloroform extraction and ethanol precipitation and subjected to EdU-labelled DNA isolation (see 'EdU-labelled DNA isolation and sequencing' below). The REPLIseq samples are listed in Supplementary Table 6.

Flow cytometry assessment of S phase entry. U2OS-CE cells were cultured with or without doxycycline for two days, and U2OS-MycER cells were cultured with or without 4OHT for three days. Cells were treated with 100 ng/ml nocodazole (Tocris, Cat. No. 1228) for 8 h to induce mitotic arrest, except for RPE1 cells, which were treated with 200 ng/ml nocodazole. Mitotic cells were isolated by shake-off, washed with PBS, released in warm medium containing 25 μM EdU (Invitrogen, Cat. No. A10044) and then collected every two hours and fixed with 90% methanol overnight. The cells were prepared for flow cytometry using the Click-it Kit (Invitrogen Cat. No. C-10420) according to the manufacturer's instructions. The genomic DNA was stained with propidium iodide (Sigma, Cat. No. 81845) in combination with RNase (Roche, Cat. No. 11119915001). EdU-DNA content profiles were then acquired by flow cytometry (Gallios, Beckman Coulter) to assess the percentage of cells that entered S phase in each condition at each time point.

EdUseq. U2OS-CE cells were cultured with or without doxycycline for two days and U2OS-MycER cells were cultured with or without 4OHT for three days before being exposed to 100 ng/ml nocodazole (Tocris, Cat. No. 1228) for 8 h to induce mitotic arrest. HeLa and RPE1 cells were treated with 100 ng/ml or 200 ng/ml nocodazole, respectively, for 8 h. Cells in mitosis were isolated by shake-off, washed with PBS, released in warm medium containing 2 mM hydroxyurea (Sigma, Cat. No. H8627) and 25 μM EdU (Invitrogen, Cat. No. A10044). When indicated, DRB (Sigma, Cat. No. D1916; 75 μM), mimosine (Sigma, Cat. No. M0253; 1 mM) or aphidicolin (Sigma, Cat. No. A0781; 1 μM) were also added to the tissue culture medium. The cells were then collected at the indicated time points and fixed with 90% methanol overnight. The EdUseq-HU samples are listed in Supplementary Table 7.

In a second series of experiments (EdUseq-noHU), after mitotic shake-off, the cells were released in medium without hydroxyurea. EdU (25 μM) was added directly to the medium or one hour before the cells were collected, as indicated. The EdUseq-noHU samples are listed in Supplementary Table 8.

In a third series of experiments (EdUseq-HU/release), after mitotic shake-off, the cells were released in medium containing hydroxyurea but not EdU. After the indicated time of incubation, hydroxyurea was removed and the cells were released in warm medium. EdU (25 μM) was added 30 min before the cells were collected, and the cells were then fixed with 90% methanol overnight. To inhibit transcription elongation, DRB was added at 75 μM to the cells for the indicated time points. The EdUseq-HU/release samples are listed in Supplementary Table 9.

For all three series of EdUseq experiments, after fixing, the cells were permeabilized with 0.2% triton-X in PBS. EdU was coupled to a cleavable biotin-azide linker (Azide-PEG(3+3)-S-S-biotin) (Jena Biosciences, Cat. No. CLK-A2112-10) using the reagents of the Click-it Kit (Invitrogen, Cat. No. C-10424). The DNA was then purified by phenol-chloroform extraction and ethanol precipitation and subjected to EdU-labelled DNA isolation (see 'EdU-labelled DNA isolation and sequencing' below).

EdU-labelled DNA isolation and sequencing. Genomic DNA was sonicated to a size range of 100–500 bp with a bioruptor sonicator (Diagenode). EdU-labelled DNA fragments were then isolated using Dynabeads MyOne streptavidin C1 (Invitrogen, Cat. No. 65001) according to the manufacturer's instructions with minor modifications. In brief, for each sample, the beads were washed three times with Binding and Washing Buffer 1× (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCL, 0.5% Tween-20) using a magnet. After washing, the beads were resuspended to twice the original volume with Binding and Washing Buffer 2×, mixed with an equal volume of sonicated EdU-labelled DNA and incubated for 15 min on a rotating wheel at room temperature. The beads were then washed three times with Binding and Washing Buffer 1× and once with TE (10 mM Tris-HCl pH 8, 1 mM EDTA) using the magnet. The EdU-labelled DNA was eluted by incubating the streptavidin beads with 2% β-mercaptoethanol (Sigma, Cat. No. M6250) for 1 h at room temperature. The eluted DNA for REPLIseq was purified by phenol–chloroform extraction followed by ethanol precipitation before being prepared for Illumina single-end sequencing. The eluted DNA for EdUseq was directly used for library preparation. The libraries were made by the Genomics Platform of the University of Geneva using the TruSeq ChIP Sample Prep Kit (Illumina, Cat. No. IP-202-1012). One hundred base pair single-end read sequencing reactions were then performed on an Illumina Hi-Seq 2500 or Illumina Hi-Seq 4000 sequencer.

In order to compare the levels of EdU incorporation among the various REPLIseq samples, the EdU-labelled genomic DNA isolated from NE and OE 2d (2 days after tetracycline withdrawal) U2OS-CE cells was spiked with a constant amount of EdU-labelled DNA prepared from mouse embryo fibroblasts (MEFs) before the EdU-labelled DNA was isolated. This permitted calibration of the amount of EdU incorporation per cell among the various samples, by dividing the number of sequencing reads of EdU-labelled human DNA by the number of reads of EdU-labelled mouse DNA and by the fraction of EdU-positive cells in that sample.

EUseq. For sequencing of newly synthesized transcripts (EUseq), cells were synchronized in mitosis and released in 2 mM hydroxyurea, as in EdUseq. For the series of experiments for which transcription elongation was inhibited, DRB was added at 75 μM to the cells for the indicated time periods. EU (5-ethynyl-uridine, Jena Biosciences, Cat. No. CLK-N002-10) was added to the cells at a concentration of 0.5 mM 30 min before the cells were collected at the indicated time points. RNA was then extracted and purified using TRIzol (Invitrogen, Cat. No. 15596) and isopropanol precipitation. Nascent RNA was biotinylated and purified using the reagents of the Click-iT Nascent RNA Capture Kit (Invitrogen, Cat. No. C-10365) according to the manufacturer's instructions, but replacing the biotin-azide from the kit with cleavable biotin-azide (Azide-PEG(3+3)-S-S-biotin) (Jena Biosciences, Cat. No. CLK-A2112-10). The EU-labelled RNA was then isolated using Dynabeads MyOne streptavidin C1 (Invitrogen, Cat. No. 65001) according to the manufacturer's instructions with minor modifications. In brief, the beads (50 μl beads per microgram RNA) were washed three times with Binding and Washing Buffer 1× (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCL, 0.5% Tween-20) followed by two 2 min washes in solution A (0.1 M NaOH, 0.05 M NaCl) and two washes in solution B (0.1 M NaCl) using a magnet. After washing, the beads were resuspended to twice the original volume with Binding and Washing Buffer 2×, and mixed with an equal volume of EU-labelled RNA. The RNA had been previously heated at 70 °C and placed back on ice to remove secondary structures. The mix was incubated for 30 min on a rotating wheel at room temperature. The beads were then washed three times with Binding and Washing Buffer 1× and once with RNase free-water using the magnet. The EU-labelled RNA was finally eluted by incubating the streptavidin beads with 2% β-mercaptoethanol (Sigma, Cat. No. M6250) for 1 h at room temperature. Sequencing libraries were prepared by the Genomics Platform of the University of Geneva using TruSeq Stranded Total RNA with Ribo-Zero Gold (Illumina, Cat. No. RS-122-2301) and omitting the ribo-depletion step. One hundred base pair single-end read sequencing reactions were performed on an Illumina Hi-Seq 2500 or Illumina Hi-Seq 4000 sequencer. The EUseq samples are listed in Supplementary Table 10.

LAM-HTGTS. Translocations were detected in NE and OE U2OS-CE cells using LAM-HTGTS as previously described[23], omitting the optional enzyme blocking step. DSBs were induced at a bait site located in an early replicating intergenic region of chromosome 9, using a guide RNA (gRNA-chr9: CACCGAGGAAACTGAGTCACAGGCT, chr9: 21685208-21685227) in combination with the Cas9 nuclease (Addgene, Cat. No. 48138, pX458, pSpCas9(BB)-2A-GFP). Cells with normal expression of cyclin E and cells in which cyclin E overexpression had been induced for three days were transfected with the Cas9:gRNA-chr9 plasmid and collected two days after transfection for the LAM-HTGTS procedure. Non-transfected cells were also collected as a negative control sample. The primers were designed as follows[23]:

Bio-primer-chr9: /5-biotin/AAGTCTCTCCAGCCAAGAACAG

I5-Nested-chr9: ACACTCTTTCCCTACACGACGCTCTTCCGATCT-BARCODE-GGAAAGGGTAGTGGGAGGTAGAAAGC

Paired-end read sequencing reactions (100 bp) were performed on an Illumina Hi-Seq 2500 sequencer. The LAM-HTGTS samples are listed in Supplementary Table 11.

Sequence alignment and calculation of sigma values. Sequence reads were aligned to the masked human genome assembly (GRCh37/hg19) using the Burrows–Wheeler aligner algorithm[31], retaining only the reads with the highest quality scores. Custom Perl scripts were then developed to process the data for analysis and visualization. First, the chromosomes were split into 10-kb bins and the sequence reads were assigned to their respective bin. Then, to correct for sequencing bias across the genome (reflecting experimental biases and differences in the number of masked base pairs per bin), the number of sequence reads per bin (SeqRpB) was normalized using the number of reads previously obtained by sequencing genomic DNA from the same cells (referred to as adjust sample and representing a total of more than 342 million reads)[14] using the formula NormSeqRpB = SeqRpB/AdjRpB, where NormSeqRpB stands for normalized sequence reads per bin and AdjRpB for adjust reads per bin. Genomic bins were retained for further analysis only if the number of adjust reads per bin was within the range of 25–10,000 (average 1,241 AdjRpB), resulting in a human genome assembly of 275,491 genomic bins corresponding to chromosomes 1–22 and chromosome X (as U2OS cells were derived from a female patient). After normalization of the sequence reads, standard deviation values were calculated for each genomic bin. This permitted the calculation of sigma (σ) values for each genomic bin according to the formula σ = (NormSeqRpB − mean background NormSeqRpB)/s.d. For all samples, the mean background NormSeqRpB value was much smaller than the peak NormSeqRpB values, but was, nevertheless, subtracted to lead to more accurate sigma value estimates. The sigma values were then used to plot the data and perform all subsequent analyses, thus, allowing comparison of samples with different background levels and providing a quick way to ascertain whether the observed peaks were statistically significant.

To calculate the genome-wide, mean background NormSeqRpB value, we used a differential function to determine the fraction of the genome with background signal (that is, the genomic bins lacking true signal); the mean NormSeqRpB value of these genomic bins corresponded to the genome-wide, background NormSeqRpB mean value. NormSeqRpB SDs were calculated for each genomic bin. First, the fraction of the genome with background signal was sorted into 200 equal subfractions, according to the number of AdjRpB, with each subfraction corresponding to a range of AdjRpB, spanning the entire possible range of 25–10,000. The NormSeqRpB values of these subfractions were then used to calculate background NormSeqRpB s.d. values for each subfraction. The s.d. values of all the subfractions were plotted against the mean AdjRpB of their corresponding subfractions, resulting in a power regression curve of the type s.d. of background NormSeqRpB of a subfraction = a × (mean AdjRpB of the subfraction)^b, where a and b are constants. For all samples, the power regression equations fit the data with coefficients of determination (R^2) greater than 0.9. The determined values of the a and b constants were then used to calculate an s.d. for each genomic bin (including the bins with true signal) from its AdjRpB. For the EUseq data, most background genomic bins had NormSeqRpB values equal to zero; thus, for these samples, we calculated relative rather than absolute s.d. and sigma values.

The EdUseq-HU data sets were plotted after subtracting the mean background NormSeqRpB value; all other data sets (including the EdUseq-HU/release and EdUseq-noHU data sets) were plotted without background subtraction.

The computer codes to perform the analyses described above are included in the Supplementary Information.

Identification and classification of replication origins from EdUseq-HU data. A peak-finding algorithm searched for local maxima. Each local maximum was then evaluated on the basis of its sigma value and shape and retained as a peak, only if its values exceeded predefined sigma and shape thresholds. One peak list comprised the peaks identified in the EdUseq-HU 4 h OE data set, while a second list comprised the peaks identified in the 14 h NE data set. The peaks that had been identified in both data sets at exactly the same genomic bin were then used to calculate an adjustment factor (AdjFactor) using the formula AdjFactor = sum of sigma values of 14 h NE shared peaks/sum of sigma values of 4 h OE shared peaks. The sigma values of all genomic bins of the 4 h OE sample were then multiplied by this adjustment factor. A new peak list was then generated by including all peaks from both data sets, irrespective of whether the peak was present in both samples or only in one sample. Peaks that mapped to adjacent genomic bins (that is, within 10 kb) in the two data sets were considered to correspond to a single origin and assigned to the genomic bin with the highest sigma value (original sigma for the 14 h NE data set and adjusted sigma for the 4 h OE data set).

For every peak in the merged peak list (irrespective of whether the peak had been identified in the NE or OE or both samples), the sigma values at its genomic position in the NE and OE samples (adjusted sigma for the OE sample) were obtained and compared. If the ratio of the OE:NE sigma values was greater than 4, then the origin was considered as oncogene-induced (Oi); otherwise, if the ratio was greater than 2, but lower than 4, the origin was considered as Oi2. All other origins were considered to be constitutive (CN). The assignment of origins into the CN, Oi2 and Oi classes facilitated comparisons among the various samples using power regression curves of the type σ of OE sample = a × (σ of NE sample)^b, where a and b are constants.

The power curve was converted to a linear regression curve: $\log_2(\sigma$ of OE sample$) = \log_2(a) + b \times \log_2(\sigma$ of NE sample$)$, to facilitate plotting of the data as scatter plots and calculation of coefficients of determination (R^2). A similar analysis was performed for cells with inducible activation of Myc and for the EUseq data.

Assignment of replication timing domains. The early, mid and late S phase REPLIseq data were calibrated by spiking the samples with a known quantity of mouse genomic EdU-labelled DNA. After assignment of the REPLIseq reads to genomic bins and adjustment of the number of reads, as described above for the EdUseq data, the number of early, mid and late S phase reads were compared for each genomic bin. If one sample (early, mid or late S) accounted for more than half of the total reads for a specific genomic bin, then that bin was assigned to the corresponding replication timing domain. The assignment of replication timing domains used for further analysis was based on the NE samples, which showed sharp REPLIseq profiles.

Assignment of genic and intergenic domains. RefSeq gene annotations were used to compile a list of all human protein-coding genes and their position in the genome. Genomic bins were defined as being purely genic, if they mapped entirely within protein-coding genes; purely intergenic, if the bins mapped entirely within intergenic sequences; or mixed, if they encompassed both genic and intergenic sequences. The analysis of the distribution of origins in the genome considered the pure intergenic and mixed genic/intergenic bins as intergenic and the pure genic bins as genic.

Determination of average EUseq and EdUseq signals along large genes. EUseq relative sigma (rσ) values were converted to a $\log_2$ scale for all the subsequent analysis described in this section. Genes over 200 kb in size were identified in the early- and mid-replicating parts of the genome using RefSeq gene annotations and the average EUseq $\log_2$(rσ) value across the lengths of these genes 14 h after mitotic exit was used to classify the genes according to their level of transcription (high, upper tercile; medium, middle tercile; and low, lower tercile). EUseq data sets corresponding to different times after exit from mitosis were then adjusted relative to each other by comparing the sum of their EUseq $\log_2$(rσ) values corresponding to the first five genomic bins of each large gene. Then, average EUseq $\log_2$(rσ) values were plotted as a function of the distance from the 5′ end of the gene. The five most 3′ genomic bins of each gene were trimmed and not included in the analysis, as in some genes EdUseq signal from origins located in intergenic bins adjacent to the gene was spilling over into the 3′ end of the gene. Average EdUseq values (linear σ) were also plotted along gene length. Sigma values of the OE samples were adjusted relative to the sigma values of the NE samples.

Calculation of fork speed. Fork speeds were calculated from EdUseq data sets of NE and OE cells treated with hydroxyurea for 14 and 10 h after mitotic exit, respectively, and EdUseq-HU/release data sets for NE and OE cells at 90 and 150 min release time points. The 10% tallest constitutive peaks in the genome were initially selected. Then, the peak finding algorithm described above was used to identify peaks in the EdUseq-HU/release 90- and 150-min data sets on either side of the origins. The positions of peaks identified with high confidence in the release data sets, were then used to calculate the distance forks travelled between the 90- and 150-min time points. The same list of origins ($n = 325$) was examined in the NE and OE samples.

Analysis of fork collapse. To study fork collapse, a set of origins was selected for which fork progression could be monitored without interference from neighbouring origins. The criterion for selection was that within 20 genomic bins of the position of the origin being examined, there was no other origin that had a sigma value equal to or greater than the sigma value of the origin being selected. Selected origins were further classified according to their transcription level (upper and lower terciles, as defined above) at the time the cells were released from the hydroxyurea arrest. The number of origins thus selected in each category was: CN-high transcription, 67; CN-low transcription, 233; Oi2-high transcription, 39; Oi2-low transcription, 55; Oi-high transcription, 57; Oi-low transcription, 47. Adjusted averages (relative to the no release data) of EdUseq and EdUseq-HU/release data sets were then calculated for each origin category. A similar analysis, using the same set of origins, was performed for cells treated with DRB.

Identification of translocations by LAM-HTGTS and mapping to Oi origins. Paired-end sequencing reads were aligned independently to the masked human genome assembly (GRCh37/hg19) using the Burrows–Wheeler aligner mem algorithm[31] and duplicate reads were filtered out. Read pairs corresponding to junctions between the bait site and another genomic region and containing the junction sequence in one of the two reads were retained. Furthermore, because DNA double-strand ends induced by fork collapse in S phase will probably be repaired by microhomology-mediated end joining, paired ends were further required to have a 2–5-base pair microhomology junction. Translocation breakpoints identified by LAM-HTGTS in OE and NE cells were then mapped to Oi origins. A subset of all Oi origins was used for this analysis, requiring that within 10 genomic bins of the position of the origin being examined, there was no other origin that had a sigma value equal to or greater than twice the sigma value of the origin being selected. Selected origins were further classified according to their transcription level (upper and lower terciles, as defined above) 14 h after the cells were released from hydroxyurea arrest. The number of selected origins were: Oi-high transcription, 108; Oi-low transcription, 62. The number of translocations mapping to each genomic bin was divided by the number of origins (Oi-high transcription or Oi-low transcription); for the translocations identified in the NE cells, the average number of translocations per bin was further adjusted by the ratio of the total number of LAM-HTGTS in the NE and OE samples to allow comparisons between the samples. A permutation analysis was performed to evaluate whether the observed differences between the number of identified translocations mapping to Oi origins in the OE and NE samples were statistically significant.

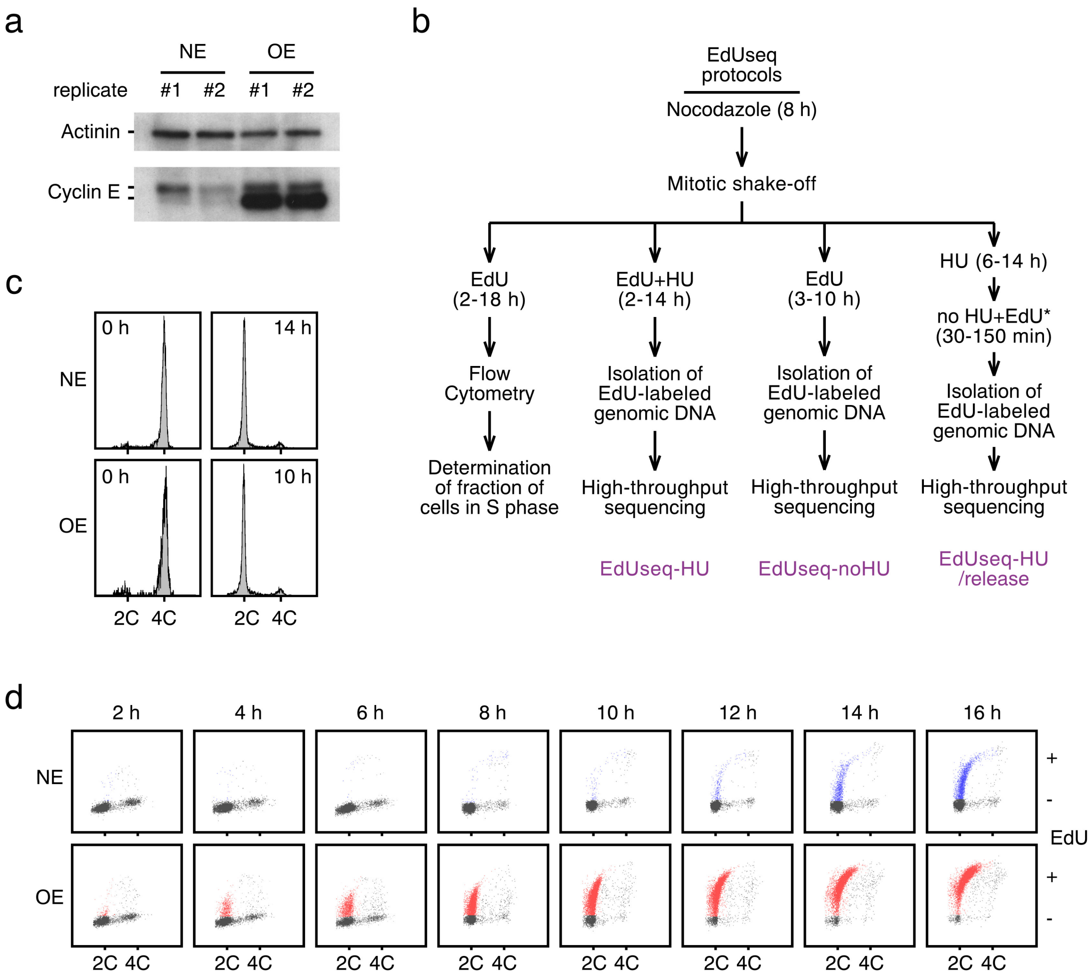
Identification of Oi replication initiation domains (OiRDs). The sigma values of the EdUseq-noHU data sets from OE cells incubated for the first 3 h after mitotic exit with EdU and from NE cells incubated for the first 10 h after mitotic exit with EdU (Fig. 1h) were converted to $\log_2$ values. A linear regression curve of the values corresponding to the genomic positions of the Oi2 origins in an OE:NE plot was then used to assign all genomic bins that had EdUseq signal above background to Oi replication initiation domain (OiRD) bins, using as criterion that the $\log_2$ OE:NE sigma values of the bin were more than 0.6 units to the left of the Oi2 curve.

Mapping of translocations and genomic rearrangement breakpoints to OiRDs. The translocation breakpoints identified by LAM-HTGTS ($n = 16,629$ and $n = 10,735$ for NE and OE cells, respectively), the breakpoints of genomic rearrangements ($n = 136$, Extended Data Table 2; derived from 81 rearrangements—for rearrangements less than 100 kb long, a single breakpoint was calculated corresponding to the centre position of the rearrangement) identified by us in the same U2OS cells overexpressing cyclin E for three weeks[14] and the breakpoints of rearrangements (deletions and amplifications, $n = 490,711$) present in a cohort of ~5,000 human cancers[4] were mapped to the Oi replication initiation domains (OiRDs). The frequency of OiRDs in the entire genome served as reference. The analysis was performed in the context of the entire genome (Fig. 4c–f) and also for only the early S replicating part of the genome (Extended Data Fig. 9a–d), where about half of the OiRDs mapped. Statistical comparisons for the LAM-HTGTS and U2OS breakpoint rearrangement data were performed using random permutations; for the TCGA data, the observed and genomic frequencies were used to calculate z-scores, from which P values were determined.

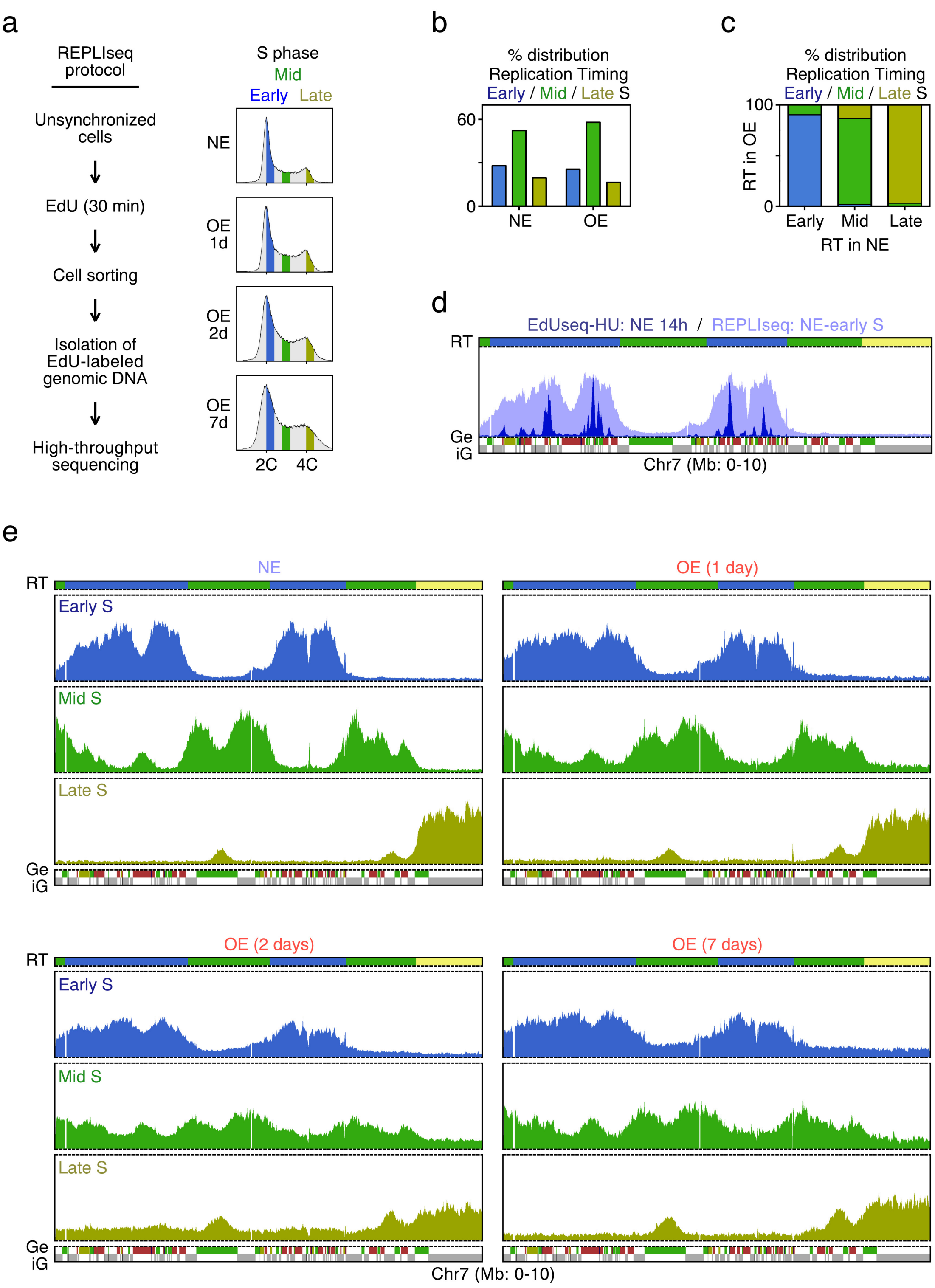
Code availability. Computer codes and data files used to process and plot the data are available as Supplementary Information. Other codes are available from the corresponding author upon request.

Data availability. The fastq sequencing data and associated information described in this study have been deposited in the Sequence Read Archive (SRA) as BioProject PRJNA397123.

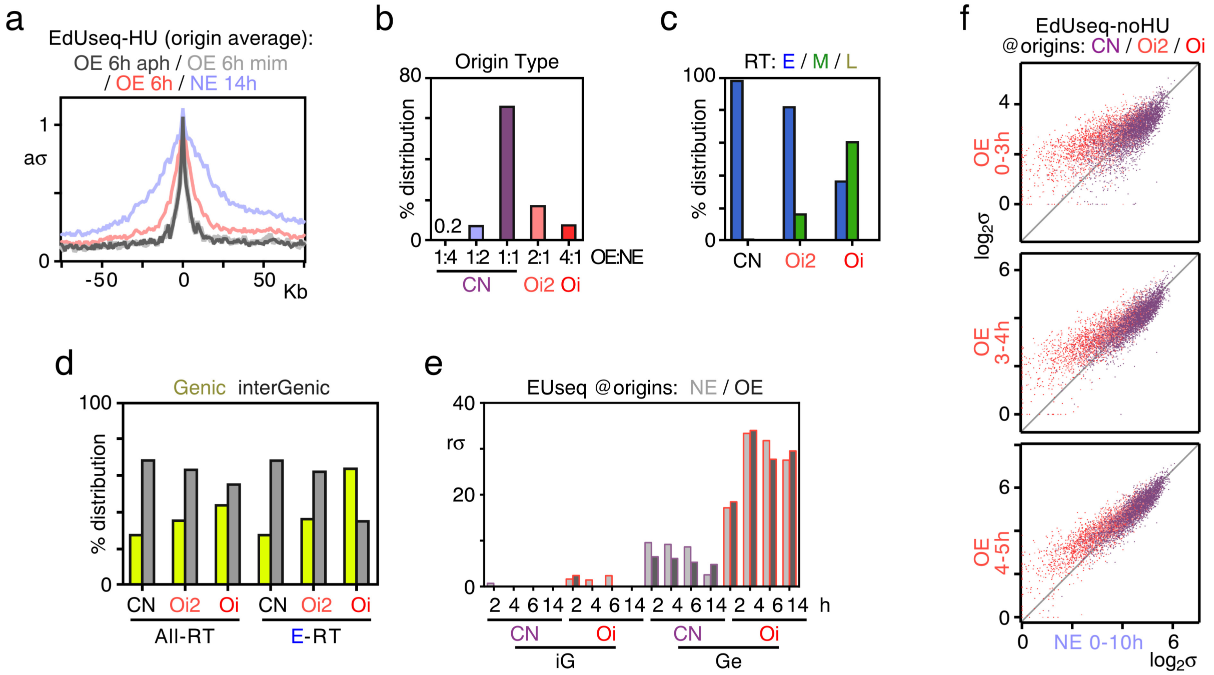
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).



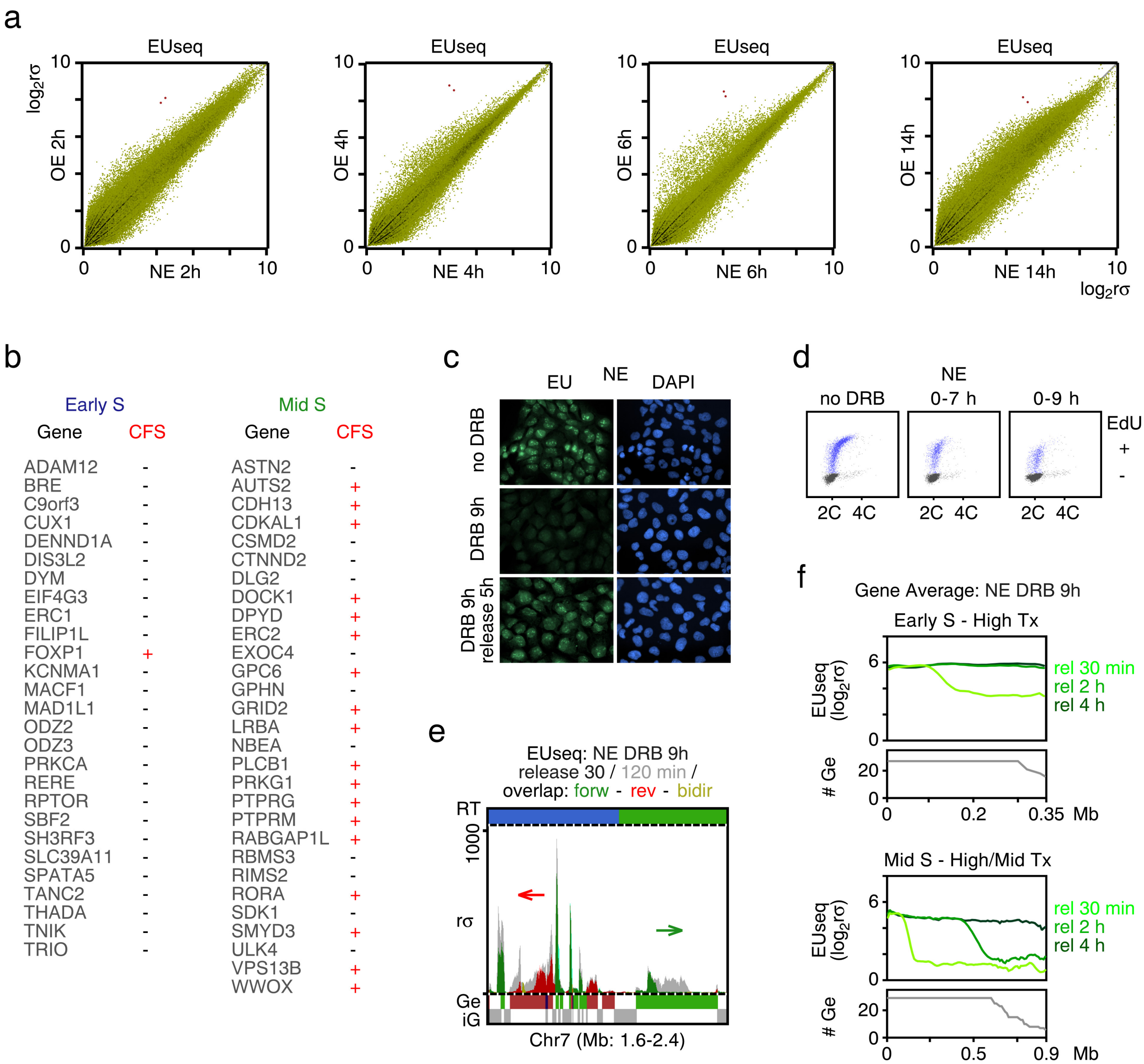
Extended Data Figure 1 | Experimental setup to study S phase entry and DNA replication initiation. **a**, Cyclin E protein levels, as determined by immunoblotting, in NE and OE cells (2.5 days after tetracycline withdrawal). Actinin serves as a loading control. This is a representative example of more than ten independent replicates. **b**, Experimental outline of the protocol used to monitor S phase entry by flow cytometry and of the EdUseq protocols: EdUseq-HU, EdUseq-noHU and EdUseq-HU/release. *EdU was added 30 min before harvesting the cells. **c**, Flow cytometry profiles of NE and OE cells after mitotic shake-off (0 h) and 14 and 10 h later, respectively, after the cells had been released in medium containing hydroxyurea and EdU. 2C and 4C, DNA content of G1 and G2 cells, respectively. This is a representative example of more than ten independent replicates. **d**, DNA content versus EdU incorporation flow cytometry plots of NE and OE cells. EdU-positive NE and OE cells were gated blue and red, respectively. 2C and 4C, DNA content of G1 and G2 cells, respectively. The gating strategy for these data is shown in Supplementary Fig. 1.



Extended Data Figure 2 | Identification of replication timing domains by REPLIseq. a, Experimental outline of the REPLIseq protocol and FACS profiles of cells with normal levels of cyclin E (NE) and cells overexpressing cyclin E (OE) for 1, 2 or 7 days (1 d, 2 d or 7 d). Cells were sorted according to DNA content into early (blue), mid (green) and late (yellow) S phase fractions. 2C and 4C, DNA content of G1 and G2 cells, respectively. **b**, Assignment of RT domains. The fractions of the genome that were replicated in early, mid or late S phase were determined on the basis of the REPLIseq profiles of the NE and 2 d OE cells. **c**, Distribution of early, mid and late replication timing bins in 2 d OE cells according to their replication timing in NE cells. **d**, Comparison of the origin firing profile determined by EdUseq-HU (Fig. 1b) and the early S replication profile determined by REPLIseq in NE cells. RT domains and Ge/iG regions are as in Fig. 1b. Bin resolution, 10 kb; ruler scale, 100 kb. **e**, REPLIseq profiles of the first 10 Mb of chromosome 7 of NE or OE 1 d, 2 d or 7 d cells. Profiles are shown separately for the cells in early, mid and late S phase. RT domains and Ge/iG regions are as in Fig. 1b. Bin resolution, 10 kb; ruler scale, 100 kb.

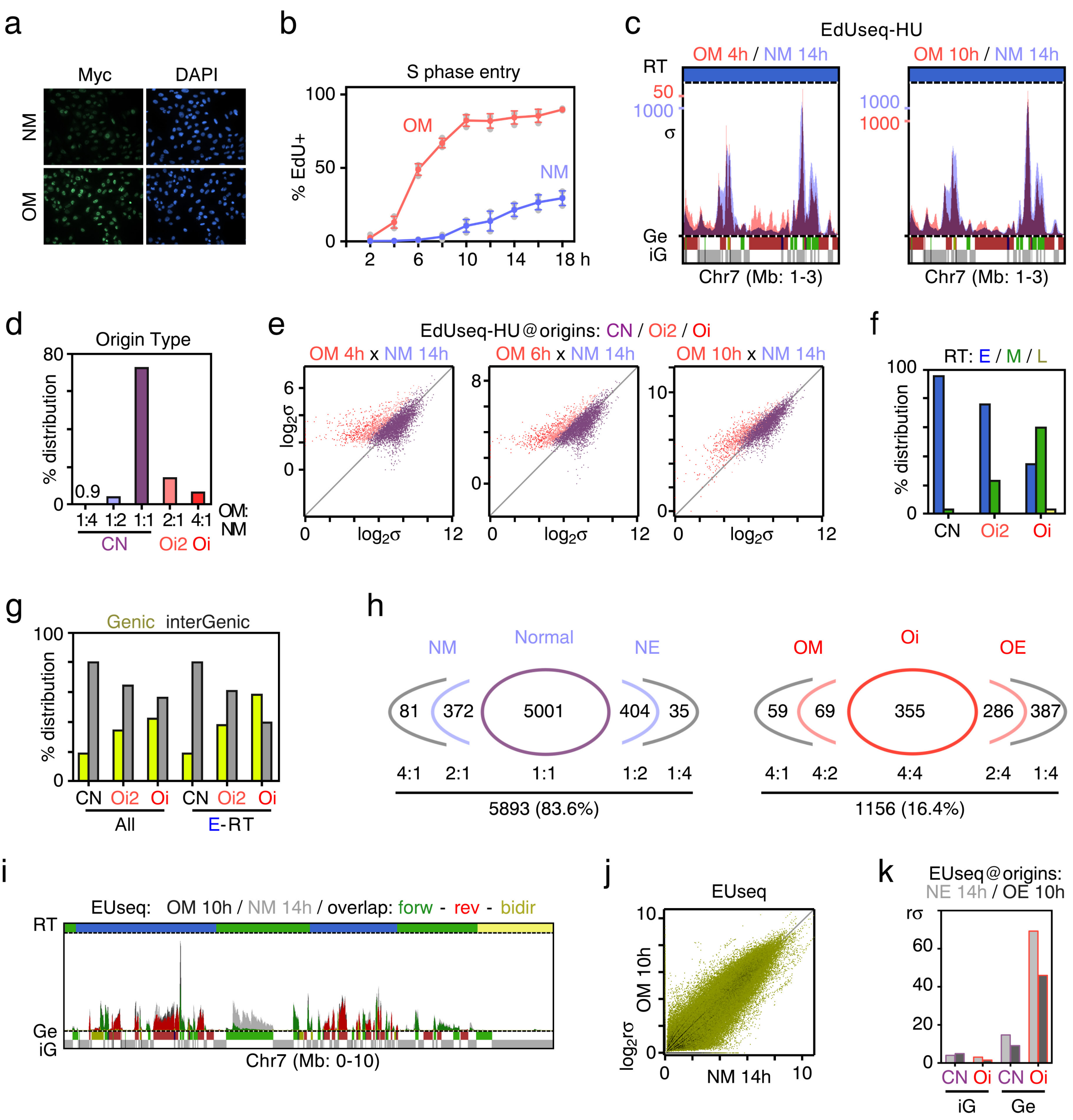


Extended Data Figure 3 | Further characterization of replication origins. a, Adjusted average sigma values at 1 kb resolution around 1,828 origins refined by performing the EdUseq-HU protocol in the additional presence of mimosine or aphidicolin (light and dark grey, respectively), compared to OE and NE cells treated only with hydroxyurea (pink and blue, respectively). **b**, Distribution of the subset of origins refined at 1 kb resolution (1,828 origins) relative to origin type, as determined by the original 10-kb resolution assignment (CN, Oi2 or Oi). **c**, Distribution of constitutive and oncogene-induced origins, refined at 1 kb resolution (1,828 origins), according to RT domains (E, early; M, mid; L, late S phase). **d**, Distribution of constitutive and oncogene-induced origins, refined at 1 kb resolution (1,828 origins), according to gene annotation in all RT domains (all-RT) or only in the early S phase RT domains (E-RT). **e**, Transcription (EUseq) levels (median) in NE (light grey) and OE (dark grey) cells at sites of constitutive and oncogene-induced origins, refined at 1 kb resolution, at various time points after mitotic shake-off. Origins mapping to genic or intergenic genomic bins were plotted separately. **f**, Scatter plots of EdUseq-noHU sigma values ($\log_2$) at all origins (CN, purple; Oi2, pink; Oi, red) at 10 kb resolution for NE versus OE cells not treated with hydroxyurea (Fig. 1h). EdU was present during the indicated time periods.

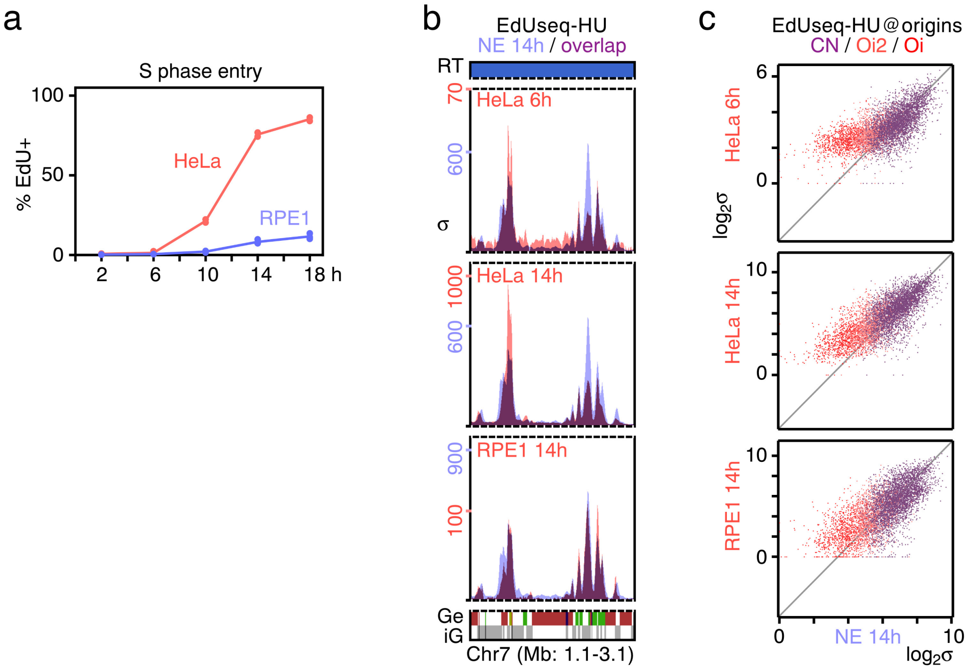


Early S Gene	CFS	Mid S Gene	CFS
ADAM12	-	ASTN2	-
BRE	-	AUTS2	+
C9orf3	-	CDH13	+
CUX1	-	CDKAL1	+
DENND1A	-	CSMD2	-
DIS3L2	-	CTNND2	-
DYM	-	DLG2	-
EIF4G3	-	DOCK1	+
ERC1	-	DPYD	+
FILIP1L	-	ERC2	+
FOXP1	+	EXOC4	-
KCNMA1	-	GPC6	+
MACF1	-	GPHN	-
MAD1L1	-	GRID2	+
ODZ2	-	LRBA	+
ODZ3	-	NBEA	-
PRKCA	-	PLCB1	+
RERE	-	PRKG1	+
RPTOR	-	PTPRG	+
SBF2	-	PTPRM	+
SH3RF3	-	RABGAP1L	+
SLC39A11	-	RBMS3	-
SPATA5	-	RIMS2	-
TANC2	-	RORA	+
THADA	-	SDK1	-
TNIK	-	SMYD3	+
TRIO	-	ULK4	-
		VPS13B	+
		WWOX	+

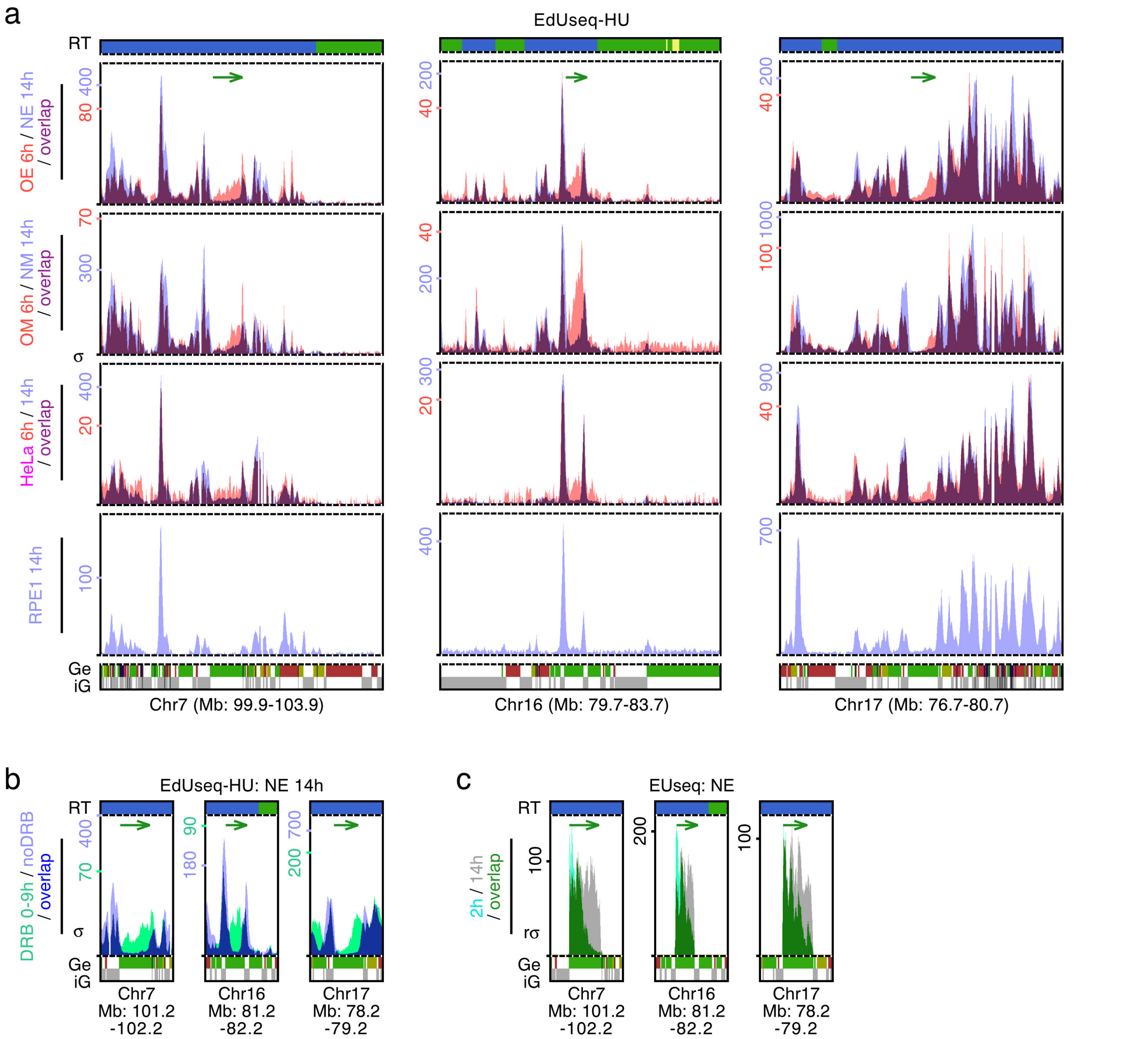
Extended Data Figure 4 | Further characterization of transcription profiles and effects of DRB. a, Genome-wide comparison (all genomic bins) of newly synthesized transcripts (EUseq) in NE versus OE cells at the indicated time points after mitotic shake-off. The two genomic bins mapping to *CCNE1* are red. **b**, List of early and mid S large genes along which replication initiation and transcription profiles were plotted in Fig. 2d. For each gene, the association with a common fragile site (CFS) is indicated[24]. **c**, Inhibition of transcription by DRB. EU incorporation was monitored by fluorescence microscopy in control (no DRB), DRB-treated (9 h), and DRB-treated (9 h) and released (5 h) cells. The nuclei of the cells were counterstained with DAPI. **d**, DNA content versus EdU incorporation flow cytometry plots of NE cells 14 h after mitotic shake-off. The cells were treated with DRB for the indicated times. EdU-positive NE cells were gated in blue. 2C and 4C, DNA content of G1 and G2 cells, respectively. **e**, Newly synthesized transcript profiles (EUseq) at a representative genomic region in NE cells treated with DRB for 9 h after mitotic shake-off and then released for 30 or 120 min (release 30 min, dark grey; release 120 min, light grey; overlap: colour; direction of transcription: green, forward (forw); red, reverse (rev); yellow, bidirectional (bidir)). The red arrow indicates the transcription of the gene harbouring oncogene-induced origins at our example locus on chromosome 7 and the green arrow indicates another large gene in this locus. RT domain and Ge/iG annotations are as in Fig. 1b. **f**, Average transcription (EUseq, $\log_2 r\sigma$) in NE cells treated with DRB for 9 h after mitotic-shake-off and then released for 30, 120 or 240 min, along the length of large genes (>0.35 Mb for early S and >0.65 Mb for mid S genes). The genes are grouped according to replication timing (RT; early, mid S) and level of transcription (high Tx, upper tercile; mid Tx, middle tercile). # Ge, number of genes averaged at each position.



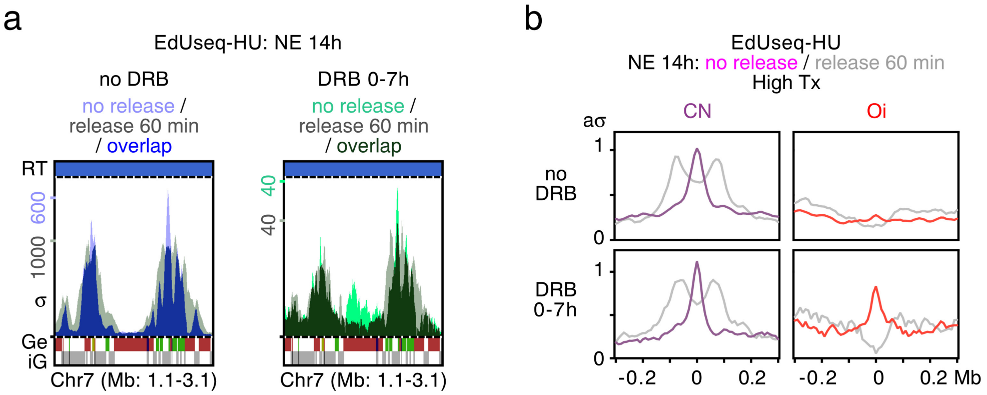
Extended Data Figure 5 | Accelerated entry into S phase and firing of novel intragenic origins upon Myc activation. a, Myc activation (3 days after adding 4-OHT), as determined by immunfluorescence, in cells with non-induced (NM) and induced (OM) Myc activity. Nuclei counterstained with DAPI. Representative images from two independent experiments are shown. **b**, Quantification of EdU-positive cells at different time points after mitotic shake-off. Means and s.d.s were calculated from three independent experiments; grey symbols, individual data points. **c**, Replication initiation (EdUseq-HU) profiles at a representative genomic region in OM and NM cells, collected at the indicated times after mitotic shake-off. Peak heights are represented as sigma values (σ). RT domains and gene annotations are as in Fig. 1b. Bin resolution, 10 kb; ruler scale, 100 kb. **d**, Classification of constitutive and oncogene-induced origins based on relative height ratios in OM versus NM cells (OM:NM). **e**, Scatter plots of EdUseq-HU σ values at origins (CN, purple; Oi2, pink; Oi, red) in NM versus OM cells at the indicated time points after mitotic shake-off. **f**, Distribution of CN, Oi2 and Oi origins in OM and NM cells according to RT domains (E, early; M, mid; L, late S phase). **g**, Distribution of CN, Oi2 and Oi origins in OM and NM cells according to gene annotation in all replication timing domains (all-RT) or only in the early S phase replicating domains (E-RT). **h**, Relative adjusted sigma ratios of replication origins identified in NE, NM, OE or OM cells. Left, number of origins identified in NE or NM cells grouped according to their relative height ratios between these two cell lines. Right, number of Oi origins identified in OE or OM cells grouped according to their level of induction relative to the NE and NM cells, respectively. **i**, Newly synthesized transcript profiles (EUseq) at a representative genomic region in OM and NM cells 10 and 14 h after mitotic shake-off, respectively (NM: light grey; OM: dark grey; overlap: green, forward (forw); red, reverse (rev); yellow, bidirectional (bidir) direction of transcription). RT domains and gene annotations are as in **c**. **j**, Genome-wide comparison (all genomic bins) of transcription in OM versus NM cells 10 and 14 h after mitotic shake-off, respectively. **k**, Median transcription (EUseq) levels in NM (light grey) and OM (dark grey) cells at constitutive and oncogene-induced origins mapping to genic (Ge) or intergenic (iG) genomic bins at 14 and 10 h after mitotic shake-off, respectively.



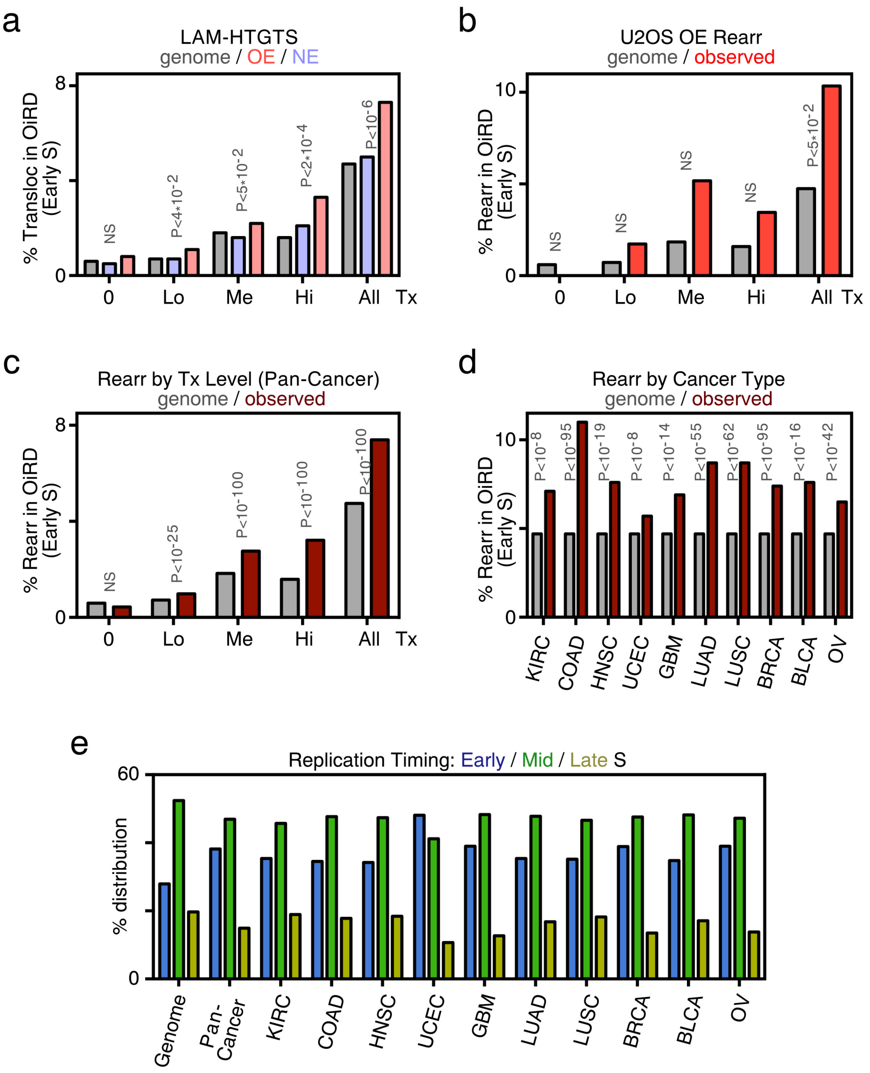
Extended Data Figure 6 | S phase entry and replication initiation profiles of HeLa and RPE1 cells. a, Percentage of EdU-positive HeLa and RPE1 cells at different time points after mitotic shake-off (0 h). Means and individual data points are shown from two independent experiments. **b**, Replication initiation (EdUseq-HU) profiles at a representative genomic region in HeLa and RPE1 cells at the indicated time points after mitotic shake-off. The profile of NE U2OS cells (blue) serves as reference. **c**, Scatter plots of EdUseq-HU σ values ($\log_2$) at all origins (CN, purple; Oi2, pink; Oi, red) in HeLa and RPE1 cells versus NE U2OS cells at the indicated time points after mitotic shake-off.



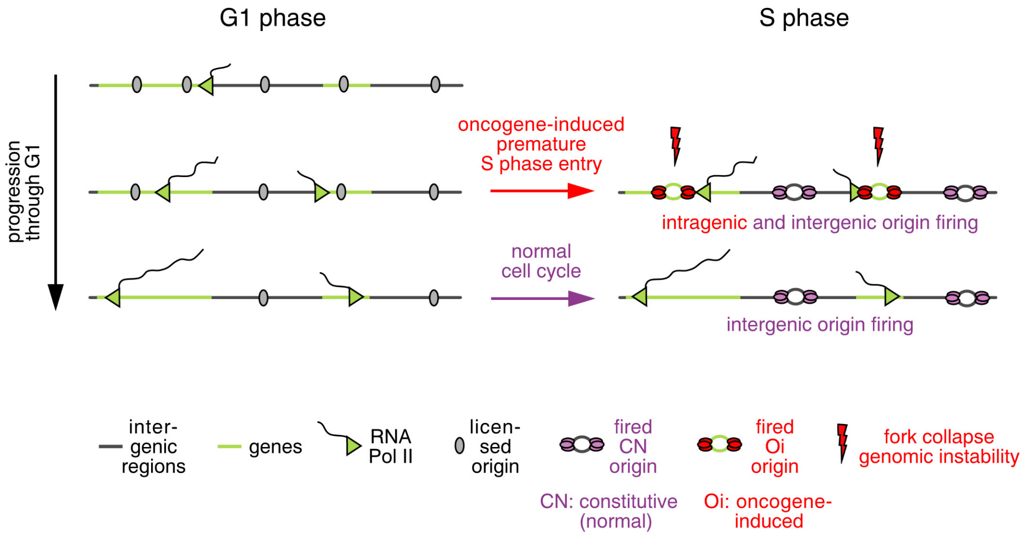
Extended Data Figure 7 | Replication initiation and transcription profiles at selected genomic loci. a, Replication initiation (EdUseq-HU) profiles at three genomic loci in different cells lines, from top to bottom: OE versus NE cells, collected 6 and 14 h after mitotic shake-off, respectively; OM versus NM cells, collected 6 and 14 h after mitotic shake-off, respectively; HeLa cells collected at 6 versus 14 h after mitotic shake-off; and RPE1 cells collected 14 h after mitotic shake-off. Peak heights are represented as sigma values. RT domains and gene annotations are as in Fig. 1b. Bin resolution, 10 kb; ruler scale, 100 kb. The green arrows indicate the direction of transcription of the example gene of each locus harbouring oncogene-induced origins. **b**, Replication initiation (EdUseq-HU) profiles of control (noDRB) and DRB-treated (0–9 h) NE cells collected 14 h after mitotic shake-off. The same genomic loci as in **a** are shown, focusing on the genes harbouring the oncogene-induced origins. RT domains and gene annotations are as in **a**. **c**, Newly synthesized transcript profiles (EUseq) of NE cells, 2 and 14 h after mitotic shake-off (2 h, light green; 14 h, grey; overlap, dark green) shown only for the example genes harbouring the oncogene-induced origins shown in **a** (green arrows). RT domains and gene annotations are as in **a**.



Extended Data Figure 8 | Fork collapse at Oi origins induced in NE cells by inhibiting transcription in early G1. a, Replication initiation (EdUseq-HU, 14 h hydroxyurea block) and fork progression (EdUseq-HU/release 60 min) profiles at a representative genomic region in NE U2OS cells treated or not with DRB during the first 7 h of G1. RT domains and gene annotations are as in Fig. 1b. Bin resolution, 10 kb; ruler scale, 100 kb. **b**, Average fork progression (no release and 60 min release) at constitutive and oncogene-induced origins located in highly transcribed regions in control and DRB treated (first 7 h of G1 phase) NE cells.



Extended Data Figure 9 | Association of Oi origins with genomic rearrangements and replication timing profiles of cancer rearrangement breakpoints. a, Mapping of translocations (Transloc; $n = 27,364$) identified by LAM-HTGTS to genomic regions replicated from Oi origins (OiRDs) with the analysis restricted to the early S replicating domains. The fraction of translocations mapping to OiRDs is shown for non-transcribed (0), low (Lo), medium (Me) and highly (Hi) transcribed genomic bins, as well as for all early S replicating bins. Statistical comparisons, using random permutation analysis, are between the NE (blue) and OE (pink) samples. The distribution of OiRDs in the genome (grey) is shown as a reference. **b**, Mapping of genomic rearrangement (Rearr; $n = 136$) breakpoints, identified previously[14] in the same OE cells, to the OiRDs, according to transcription levels, as in **a**, with the analysis restricted to the early S replicating domains. Statistical comparisons, using random permutation analysis, are between observed (red) and genomic (grey) frequencies. NS, not significant. **c**, Mapping of genomic rearrangement (Rearr; $n = 490,711$) breakpoints from a TCGA pan-cancer data set[4] to the OiRDs, according to transcription levels, as in **a**, with the analysis restricted to the early S replicating domains. Statistical comparisons, using *z*-scores, are between observed and genomic frequencies. **d**, Mapping of genomic rearrangement (Rearr; $n = 490,711$) breakpoints in common cancer types from a TCGA pan-cancer data set[4] to the OiRDs, with the analysis restricted to the early S replicating domains. KIRC, kidney renal cell; COAD, colon adenocarcinoma; HNSC, head and neck squamous cell; UCEC, uterine cervix; GBM, glioblastoma multiformae; LUAD, lung adenocarcinoma; LUSC, lung squamous cell; BRCA, breast; BLCA, bladder; OV, ovary. Statistical comparisons, using *z*-scores, are between observed (red) and genomic (grey) frequencies. **e**, Distribution of cancer rearrangement breakpoints[4] according to the replication timing data obtained from the REPLIseq experiment shown in Extended Data Fig. 2.



Extended Data Figure 10 | Proposed mechanism for oncogene-induced DNA replication stress. During the length of a normal G1 phase, transcription progressively inactivates intragenic origins, such that upon S phase entry origin firing is restricted to intergenic domains. Following oncogene activation, cells enter prematurely into S phase, before the inactivation of all intragenic origins. This results in bidirectional forks within highly transcribed genes, leading to conflicts between the replication and transcription machineries, fork collapse, DNA DSBs and genomic instability.

doi:10.1038/nature25502

The cryo-electron microscopy structure of huntingtin

Qiang Guo[1]*, Bin Huang[2]*, Jingdong Cheng[3], Manuel Seefelder[2], Tatjana Engler[2], Günter Pfeifer[1], Patrick Oeckl[4], Markus Otto[4], Franziska Moser[5], Melanie Maurer[5], Alexander Pautsch[5], Wolfgang Baumeister[1], Rubén Fernández-Busnadiego[1] &
Stefan Kochanek[2]

Huntingtin (HTT) is a large (348 kDa) protein that is essential for embryonic development and is involved in diverse cellular activities such as vesicular transport, endocytosis, autophagy and the regulation of transcription[1,2]. Although an integrative understanding of the biological functions of HTT is lacking, the large number of identified HTT interactors suggests that it serves as a protein–protein interaction hub[1,3,4]. Furthermore, Huntington's disease is caused by a mutation in the *HTT* gene, resulting in a pathogenic expansion of a polyglutamine repeat at the amino terminus of HTT[5,6]. However, only limited structural information regarding HTT is currently available. Here we use cryo-electron microscopy to determine the structure of full-length human HTT in a complex with HTT-associated protein 40 (HAP40; encoded by three F8A genes in humans)[7] to an overall resolution of 4 Å. HTT is largely α-helical and consists of three major domains. The amino- and carboxy-terminal domains contain multiple HEAT (huntingtin, elongation factor 3, protein phosphatase 2A and lipid kinase TOR) repeats arranged in a solenoid fashion. These domains are connected by a smaller bridge domain containing different types of tandem repeats. HAP40 is also largely α-helical and has a tetratricopeptide repeat-like organization. HAP40 binds in a cleft and contacts the three HTT domains by hydrophobic and electrostatic interactions, thereby stabilizing the conformation of HTT. These data rationalize previous biochemical results and pave the way for improved understanding of the diverse cellular functions of HTT.

Computational and biochemical studies of HTT have predicted a variable number of HEAT repeats interspersed by unstructured regions[8–12]. However, attempts to determine the structure of HTT at high resolution have been hindered by its flexibility[13–15]. Most structural studies have focused on an N-terminal fragment corresponding to the first exon of the *HTT* gene, and the majority of the protein (more than 97% of its amino acid length) remains largely uncharted[14]. To overcome this hurdle we searched for interaction partners that could stabilize the structure of HTT. A first screen using polyglutamine-expanded (46 glutamine) full-length human HTT (46QHTT) expressed at low levels in HEK293 cells identified abundant binding with HAP40 (Fig. 1a), which has been previously reported to interact with HTT[7] and to recruit HTT to early endosomes[16]. Although a complex of HTT and HAP40 could not be reconstituted from the individual proteins *in vitro*, the complex was purified at high yield from human cells co-expressing both wild-type full-length human HTT (17QHTT) and HAP40 (Fig. 1b). Whereas HTT alone formed oligomers and tended to aggregate[17], the HTT–HAP40 complex eluted as a symmetric narrow peak during size-exclusion chromatography (Fig. 1c). Ultracentrifugation analysis consistently indicated that the HTT–HAP40 complex was more conformationally homogeneous than HTT alone (Extended Data Fig. 1). The HTT–HAP40 complex, but not its isolated components, showed a sharp, strong unfolding transition in differential scanning

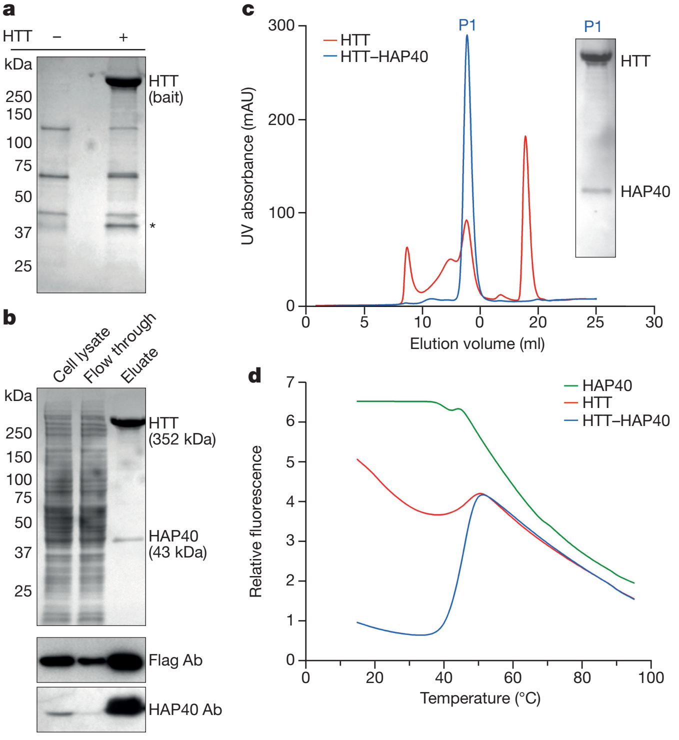


Figure 1 | Purification of the HTT–HAP40 complex. a, Identification of HAP40 as a major HTT interactor in HEK293 cells expressing low levels of Flag-tagged 46QHTT[17] (+, right lane) or not expressing Flag-tagged HTT (−, left lane). Coomassie-stained gel after PAGE following Flag-affinity purification. The band indicated by an asterisk was identified as HAP40 by mass spectrometry and western blot. **b**, Purification of the HTT–HAP40 complex from HEK293-based cells expressing Flag-tagged 17QHTT and Strep-tagged HAP40. Cleared lysates were incubated with Strep-Tactin beads and washed with desthiobiotin to elute bound proteins. Top, Coomassie staining; bottom, western blot. **c**, Elution profile of HTT alone (red) versus the HTT–HAP40 complex (blue). mAU, milli-absorbance units. Inset, Coomassie-staining of the P1 peak of the elution profile of the HTT–HAP40 complex. **d**, Thermal unfolding and complex stabilization. Melting curves of HTT, HAP40 and the HTT–HAP40 complex obtained by differential scanning fluorimetry. Data are representative of two (**a**, **d**) or three (**b**, **c**) independent experiments. For gel source images see Supplementary Fig. 1. Source Data for the differential scanning fluorimetry experiment are available online.

[1]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany. [2]Department of Gene Therapy, Ulm University, 89081 Ulm, Germany. [3]Gene Center, Department of Biochemistry and Center for Integrated Protein Science Munich, Ludwig-Maximilians University, 81377 Munich, Germany. [4]Department of Neurology, Ulm University, 89081 Ulm, Germany. [5]Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach an der Riß, Germany.
*These authors contributed equally to this work.

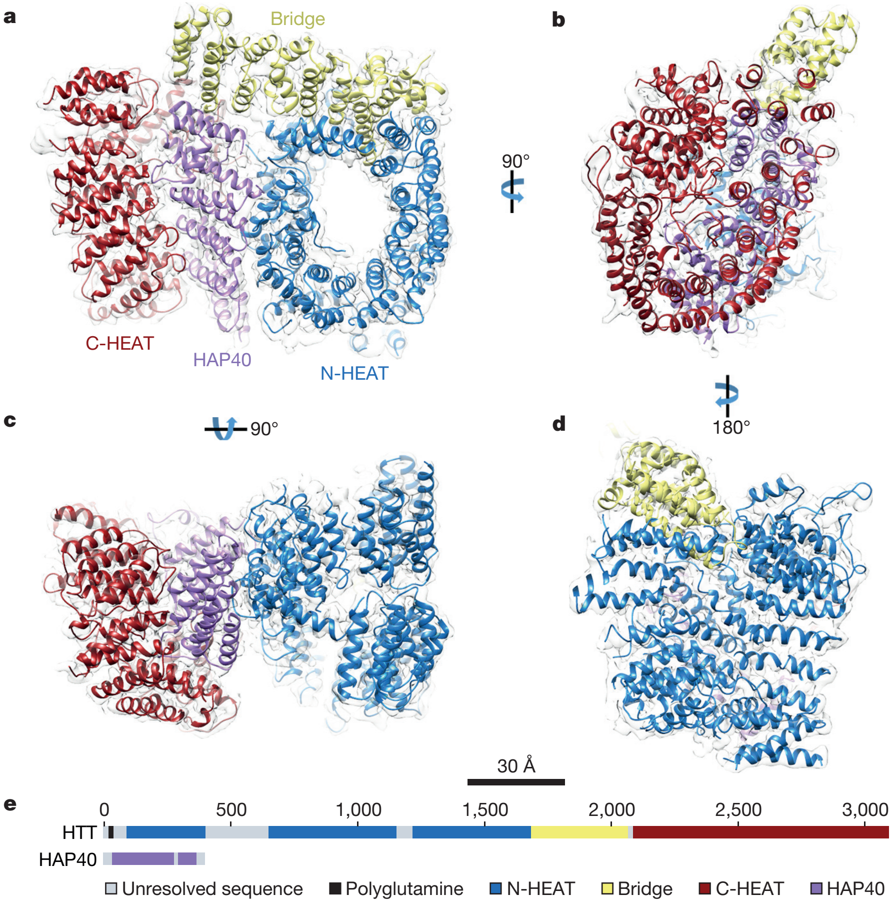


Figure 2 | Architecture of the HTT–HAP40 complex. a–d, The reconstructed density map filtered according to local resolution is shown as a translucent surface. The atomic model is superimposed in ribbon representation, with domains colour-coded as follows: HTT N-HEAT domain, blue; HTT bridge domain, yellow; HTT C-HEAT domain, maroon; HAP40, purple. **a**–**d** show different views of the complex as indicated. **e**, Schematic of the domain organization of HTT and HAP40.

fluorimetry assays (Fig. 1d), confirming that the complex was stable and amenable to structural studies[18].

In the absence of an interactor, the conformational heterogeneity of HTT prevented high-resolution cryo-electron microscopy (cryo-EM) analysis. By contrast, the HTT–HAP40 complex was well defined and yielded a globular structure measuring approximately 120 × 80 × 100 Å (Fig. 2, Extended Data Fig. 2a, b), which was to some extent reminiscent of a published negative-stain structure of HTT[9]. The global resolution of the map was 4 Å (Extended Data Fig. 2c, d), sufficient to build a *de novo* atomic model using energy minimization with well-resolved large side chains as landmarks (Fig. 2, Extended Data Figs 2e, 3, Extended Data Table 1). For both HTT and HAP40, all secondary-structure elements resolved in the model corresponded to α-helices (Extended Data Fig. 4), in agreement with computational predictions using PSIPRED[19] (Extended Data Fig. 5). For HTT, 72% of the helices were arranged in HEAT or other tandem repeats. On the other hand, most of the regions not resolved in the map were predicted to be unstructured. Notably, no density was observed for the HTT exon 1 fragment (residues 1–90; 17QHTT is used for amino acid numbering throughout the text) even at very low thresholds, indicating that this region of the protein is extremely flexible. Thus, polyglutamine length may have limited influence on the overall architecture of the HTT–HAP40 complex.

The domain organization of HTT has been controversial[1,8–12]. Our data show that HTT consists of three domains: N- and C-terminal domains containing multiple HEAT repeats (hereafter N-HEAT and C-HEAT) linked by a smaller bridge domain (Fig. 2). N-HEAT (residues 91–1,684) forms a typical α-solenoid, comprising 21 HEAT repeats arranged as a one-and-a-half-turn right-handed superhelix, the concave face of which defines an arch of approximately 80 Å in diameter (Fig. 3a). Two putative membrane-binding regions have been identified in HTT, both within N-HEAT: an exon 1 fragment, especially residues 1–17, which may form an amphipathic helix[20], and a larger region at residues 168–366, which contains a functionally important palmitoylation site at C208[21,22]. Although the N terminus corresponding to exon 1 is not visible in our structure, N-HEAT repeats 2–4 (residues 160–275) form a positively charged region at the second putative membrane-binding region in the N-HEAT convex surface (Fig. 4a). However, a previously reported putative amphipathic helix (residues 223–240)[22] faces the inner concave side of N-HEAT, with limited accessibility to membrane interactions.

Consistent with our computational predictions (Extended Data Fig. 5) and previous studies[1,14], N-HEAT accommodates a large disordered insertion (residues 400–674) between N-HEAT repeats 6 and 7 (Fig. 3a). The insertion projects outwards without interrupting the interactions between HEAT repeats and makes this region accessible to protease action. Multiple proteolytic cleavage sites have been mapped to this region[1,14,23], but the continuous packing of N-HEAT repeats 6 and 7 makes it unlikely that such cleavage events would result in an easy release of N-terminal fragments, consistent with previous reports[15,24]. This insertion also harbours multiple phosphorylation sites that may modulate protein–protein interactions and proteolytic accessibility[1,25–27], perhaps by regulating the interaction of this insertion with the positively charged region of N-HEAT. Most other reported post-translational modifications of HTT are located in presumably unstructured regions that are not resolved in our map (Extended Data Fig. 4), including protease cleavage sites that release N-terminal HTT fragments[1,23,28].

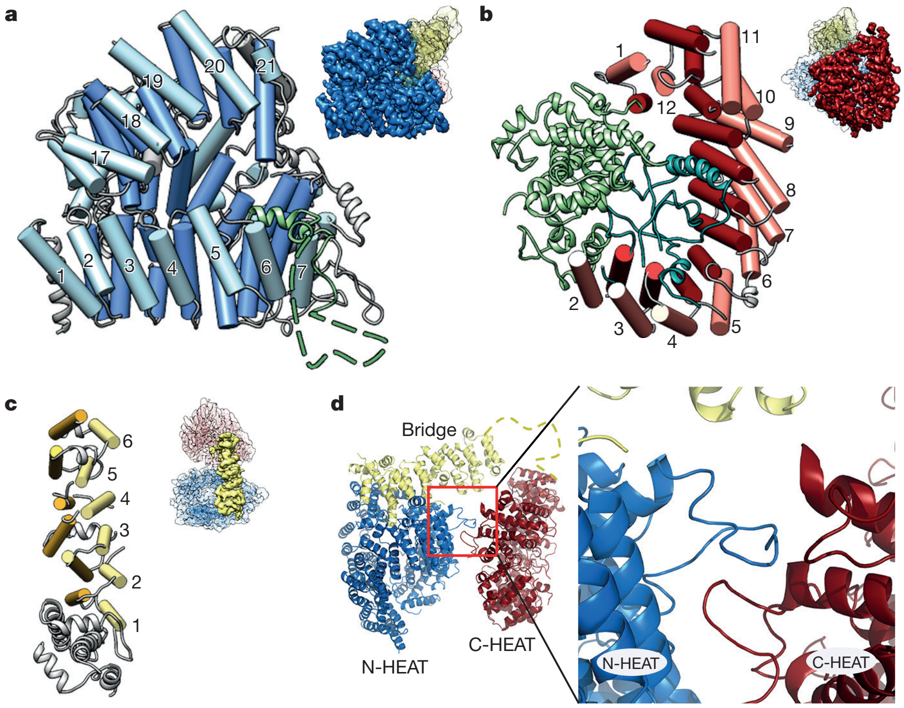


Figure 3 | Structure of HTT domains. a, N-HEAT domain. The insertion between N-HEAT repeats 6 and 7 is shown in green, with the unresolved sequence as a dashed line. **b**, C-HEAT domain, with the insertions between HEAT repeats shown in green (between C-HEAT repeats 1 and 2) and teal (between C-HEAT repeats 2 and 3). **c**, Bridge domain. In **a**–**c**, helices forming part of tandem repeats are shown as rods in similar colours, other helices are shown as ribbons and the region of HTT shown in each panel is highlighted in the HTT–HAP40 map in the upper right corner of each panel (colour code matches Fig. 2). Helical repeats are numbered from N to C terminus. **d**, Reverse view of HTT highlighting the region of interaction (expanded view) between the loops of the N- and C-HEAT domains. The unresolved sequence at the C terminus of the bridge domain is shown as a yellow dashed line.

C-HEAT (residues 2,092–3,098) comprises 12 HEAT repeats forming an elliptical ring of approximately 80 × 30 Å (Fig. 3b) with repeats 1 and 12 interacting to close the ring. Unlike canonical HEAT repeats, in which the first helix is exposed to the convex surface of the domain, the first helix of C-HEAT repeat 12 faces the concave surface of C-HEAT. The repeats are interrupted by two insertions. Insertion 1 (residues 2,121–2,456) consists of 12 helical segments that separate C-HEAT repeats 1 and 2. On the other hand, insertion 2 (residues 2,510–2,663) is mostly unstructured and does not interfere with the interaction between C-HEAT repeats 2 and 3. Both insertions are harboured in the concave surface of C-HEAT, potentially shielding this region from protein–protein interactions. By contrast, both the convex and concave surfaces of N-HEAT are accessible in the structure (Fig. 3a) and could thus act as cargo-binding sites. This may explain why most of the known binding sites of HTT interactors have been mapped to its N terminus[1,3,29].

N-HEAT and C-HEAT are stacked approximately vertically and connected by the bridge domain (residues 1,685–2,091, Fig. 3c). This domain contains six tandem α-helical repeats, of which repeats 3, 4 and 6 are armadillo-like. The repeat region is flanked by five non-repeat

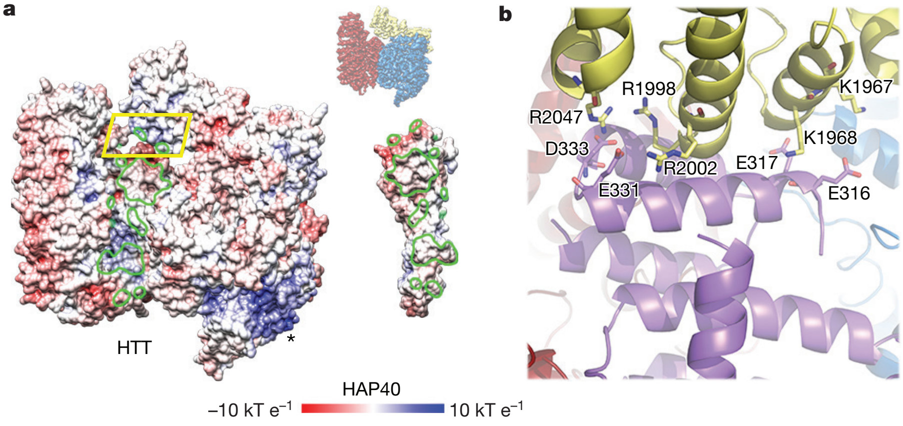


Figure 4 | Structural basis of the HTT–HAP40 interaction. a, 'Open book' view of HTT and HAP40 to show the interaction surfaces of both proteins, also displaying electrostatic surface potentials. HTT–HAP40 contact areas are outlined in green. The asterisk marks the positively charged surface formed by N-HEAT repeats 2–4. The region outlined in yellow is shown in more detail in **b**. At the top right, the orientation of the HTT–HAP40 complex with respect to Fig. 2 is indicated. **b**, Detailed view of the outlined region in **a** showing the interface between the C terminus of HAP40 and the bridge domain of HTT. Residues involved in electrostatic interactions are displayed as sticks; colour coding matches Fig. 2.

helices and a flexible C terminus (residues 2,062–2,092), which is unresolved. Besides this flexible linkage, N-HEAT and C-HEAT are connected only weakly via loop interactions (Fig. 3d), explaining the highly dynamic structure of HTT in the absence of interaction partners such as HAP40.

HAP40 binds within the cleft defined by the two HEAT domains and the bridge domain, thereby stabilizing the observed HTT conformation. HAP40 consists of 14 α-helices arranged in tetratricopeptide repeat-like tandem repeats (Fig. 2). Within the complex, HTT and HAP40 share large interfaces with mainly hydrophobic interactions (Fig. 4a). This finding is consistent with our differential scanning fluorimetry data, which suggest that the exposure of hydrophobic areas is considerably reduced in the HTT–HAP40 complex (Fig. 1d). The C terminus of HAP40 contains four negatively charged residues that interact with a positively charged patch on the bridge domain of HTT (Fig. 4b). By contrast, the N terminus of HAP40 is mostly solvent exposed, and consequently helix 1 is not well resolved in our map. Similarly, a central region of HAP40 (residues 217–258) was not visible, consistent with biochemical experiments showing that this region was not required for HTT binding (Extended Data Fig. 6).

Although HTT is highly conserved from sea urchins to humans[12] (Extended Data Fig. 7), the HTT orthologue in *Drosophila melanogaster* bears little resemblance to human HTT. No HAP40 homologue appears to be present in *D. melanogaster*, suggesting that these two proteins may have co-evolved. Many HTT interactors bind the N terminus of HTT, whereas the binding of HAP40 to HTT requires the coordination of all HTT domains (Figs 2, 4). This explains why HAP40 has been identified as an HTT interactor by previous studies that used full-length HTT as bait[4,7], but not by others that only used fragments of HTT[3]. It is also possible that other proteins bind to HTT at a similar location. Taken together, our data resolve long-standing speculations regarding the architecture of HTT, strongly support the concept that HTT serves as a multivalent interaction hub[1] and invite future structure-guided studies of the mechanisms by which HTT coordinates its diverse activities.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 November 2017; accepted 8 January 2018.
Published online 21 February 2018.

1. Saudou, F. & Humbert, S. The biology of huntingtin. *Neuron* **89,** 910–926 (2016).
2. Zuccato, C. & Cattaneo, E. in *Huntington's Disease* 4th edn (eds Bates, G. *et al.*) Ch. 11 (Oxford Univ. Press, 2014).
3. Kaltenbach, L. S. *et al.* Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet.* **3,** e82 (2007).
4. Shirasaki, D. I. *et al.* Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* **75,** 41–57 (2012).
5. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983 (1993).
6. Finkbeiner, S. Huntington's disease. *Cold Spring Harb. Perspect. Biol.* **3,** a007476 (2011).
7. Peters, M. F. & Ross, C. A. Isolation of a 40-kDa Huntingtin-associated protein. *J. Biol. Chem.* **276,** 3188–3194 (2001).
8. Palidwor, G. A. *et al.* Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput. Biol.* **5,** e1000304 (2009).
9. Vijayvargia, R. *et al.* Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *eLife* **5,** e11184 (2016).
10. Andrade, M. A. & Bork, P. HEAT repeats in the Huntington's disease protein. *Nat. Genet.* **11,** 115–116 (1995).
11. Takano, H. & Gusella, J. F. The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC Neurosci.* **3,** 15 (2002).
12. Tartari, M. *et al.* Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. *Mol. Biol. Evol.* **25,** 330–338 (2008).
13. Seong, I. S. *et al.* Huntingtin facilitates polycomb repressive complex 2. *Hum. Mol. Genet.* **19,** 573–583 (2010).
14. Wetzel, R. & Mishra, R. in *Huntington's Disease* 4th edn (eds Bates, G. *et al.*) Ch. 12 (Oxford Univ. Press, 2014).
15. Li, W., Serpell, L. C., Carter, W. J., Rubinsztein, D. C. & Huntington, J. A. Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J. Biol. Chem.* **281,** 15916–15922 (2006).
16. Pal, A., Severin, F., Lommer, B., Shevchenko, A. & Zerial, M. Huntingtin–HAP40 complex is a novel Rab5 effector that regulates early endosome motility and is up-regulated in Huntington's disease. *J. Cell Biol.* **172,** 605–618 (2006).
17. Huang, B. *et al.* Scalable production in human cells and biochemical characterization of full-length normal and mutant huntingtin. *PLoS ONE* **10,** e0121055 (2015).
18. Chari, A. *et al.* ProteoPlex: stability optimization of macromolecular complexes by sparse-matrix screening of chemical space. *Nat. Methods* **12,** 859–865 (2015).
19. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res.* **41,** W349–W357 (2013).
20. Arndt, J. R., Chaibva, M. & Legleiter, J. The emerging role of the first 17 amino acids of huntingtin in Huntington's disease. *Biomol. Concepts* **6,** 33–46 (2015).
21. Yanai, A. *et al.* Palmitoylation of huntingtin by HIP14 is essential for its trafficking and function. *Nat. Neurosci.* **9,** 824–831 (2006).
22. Kegel, K. B. *et al.* Huntingtin associates with acidic phospholipids at the plasma membrane. *J. Biol. Chem.* **280,** 36464–36473 (2005).
23. Wellington, C. L. *et al.* Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J. Biol. Chem.* **273,** 9158–9167 (1998).
24. El-Daher, M. T. *et al.* Huntingtin proteolysis releases non-polyQ fragments that cause toxicity through dynamin 1 dysregulation. *EMBO J.* **34,** 2255–2271 (2015).
25. Luo, S., Vacher, C., Davies, J. E. & Rubinsztein, D. C. Cdk5 phosphorylation of huntingtin reduces its cleavage by caspases: implications for mutant huntingtin toxicity. *J. Cell Biol.* **169,** 647–656 (2005).
26. Schilling, B. *et al.* Huntingtin phosphorylation sites mapped by mass spectrometry. Modulation of cleavage and toxicity. *J. Biol. Chem.* **281,** 23686–23697 (2006).
27. Ratovitski, T. *et al.* Post-translational modifications (PTMs), identified on endogenous huntingtin, cluster within proteolytic domains between HEAT repeats. *J. Proteome Res.* **16,** 2692–2708 (2017).
28. Ratovitski, T. *et al.* Mutant huntingtin N-terminal fragments of specific size mediate aggregation and toxicity in neuronal cells. *J. Biol. Chem.* **284,** 10855–10867 (2009).
29. Gusella, J. F. & MacDonald, M. E. Huntingtin: a single bait hooks many species. *Curr. Opin. Neurobiol.* **8,** 425–430 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Plitzko for electron microscopy support, F. Beck for help with image processing, E. Conti, H. Kiefer, B. Landwehrmeyer, K. Lindenberg, P. Mittl and L. Toledo-Sherman for discussions and A. Bracher, M. Hipp and E. Sakata for the critical reading of the manuscript. This work has been funded by the CHDI Foundation, the German Federal Ministry of Education and Research (FTLDc 01GI1007A), the German Research Foundation (SFB1279) and the European Commission (grant FP7 GA ERC-2012-SyG_318987–ToPAG). Q.G. is the recipient of postdoctoral fellowships from EMBO (EMBO ALTF 73-2015) and the Alexander von Humboldt Foundation.

Author Contributions Q.G., B.H., P.O., A.P., W.B., R.F.-B. and S.K. designed experiments. F.M., M.M. and A.P. performed differential scanning fluorimetry studies. P.O. performed mass spectrometry analyses. B.H. prepared the HTT–HAP40 samples for cryo-EM and performed the majority of the biochemical work together with T.E. M.S. performed the experiments with truncated HAP40 constructs. Q.G. performed the majority of the cryo-EM work. Q.G. and G.P. optimized sample conditions for cryo-EM. J.C. built the atomic model. Q.G., J.C. and R.F.-B. analysed the structure. Q.G., B.H., J.C., M.S., P.O., M.O., A.P., R.F.-B. and S.K. analysed the data. Q.G. and B.H. prepared the figures. Q.G., B.H., W.B., R.F.-B. and S.K. wrote the manuscript. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to W.B. (baumeist@biochem.mpg.de), R.F.-B. (ruben@biochem.mpg.de) and S.K. (stefan.kochanek@uni-ulm.de).

Reviewer Information *Nature* thanks S. Scheres and R. Wetzel for their contribution to the peer review of this work.

METHODS

Antibodies. The following antibodies were used: anti-Flag M2 (Sigma), anti-HAP40 (Santa Cruz SC-69489), anti-Strep (IBA 2-1507-001) and anti-HTT (Millipore MAB2166).

Identification of HTT-interacting proteins. HEK293-based C2.6 cells[17] (2×10^8) expressing Flag-tagged full-length polyglutamine-expanded 46QHTT at low levels were collected, lysed with 25 mM Tris, 150 mM NaCl, 0.5% Tween 20, 1× protease inhibitor (Roche), pH 7.4 and then centrifuged (20,000*g*, 1 h). The supernatant was incubated with Flag beads at 4 °C for 2 h and was then washed three times with 25 mM Tris, 150 mM NaCl, 0.02% Tween 20, pH 7.4. Proteins bound to the Flag beads were eluted with 100 mM glycine, 150 mM NaCl, 0.02% Tween, pH 3.5 and immediately neutralized with 1 M Tris (pH 8.0). The eluted proteins were concentrated and analysed by SDS–PAGE and Coomassie staining. To identify potential interactors of HTT the lanes were excised and the proteins were in-gel digested using trypsin, and then analysed by nano liquid chromatography (C18, 500 × 0.075 mm, 2 μm column, Thermo Fisher Scientific) and tandem mass spectrometry (QExactive, Thermo Fisher Scientific) in data-dependent acquisition mode (Top12). Proteins were identified using Proteome Discoverer 1.4 (Thermo Fisher Scientific) with a peptide false discovery rate ≤ 0.01, and enrichment analysis was performed with Perseus 1.4.1.3 using MS1 peak area for quantification.

The identity of Flag affinity-purified proteins was confirmed by western blot analysis with anti-HTT and anti-HAP40 antibodies after SDS–PAGE.

Generation of a stable human cell line co-expressing 17QHTT and HAP40. B1.21 cells[17] are based on HEK293 cells and express full-length wild-type 17QHTT upon induction with doxycycline. The expression plasmid pBSK/2-CMV-HAP40-TS was constructed to express the human HAP40 protein (NCBI RefSeq NP_036283.2) with a C-terminal Twin-Strep-tag under the control of the hCMV promoter. B1.21 cells were co-transfected with this plasmid together with a puromycin resistance gene. The resulting stable cell line (B1.21-HAP40TS), which expresses HAP40 at high constitutive levels and HTT upon induction with doxycycline, was used for the purification of the HTT–HAP40 complex. Generated cell lines tested negative for mycoplasma by PCR. Cell lines were authenticated by inducibility of HTT expression with doxycycline and western blot analysis.

Purification of HTT, HAP40 and the HTT–HAP40 complex. The purification of HTT alone has been described[17]. For purification of the HTT–HAP40 complex, 2×10^8 B1.21-HAP40TS cells were collected 72 h after induction with doxycycline by centrifugation at 400*g* for 10 min. Cells were lysed with 25 mM HEPES, 300 mM NaCl, 0.5% Tween 20, protease inhibitor, pH 8.0 by rotation at 4 °C for 30 min followed by centrifugation of the cell lysate at 30,000*g* and clearance by filtration through a 0.2-μm filter. The filtrate was incubated with Strep-Tactin beads (Qiagen) for 2–3 h at 4 °C. After washing three times with 25 mM HEPES, 300 mM NaCl, 0.02% Tween 20, pH 8.0, bound proteins were eluted with 25 mM HEPES, 300 mM NaCl, 0.02% Tween 20, 2.5 mM desthiobiotin, pH 8.0. The eluate was concentrated using Amicon filters.

The HTT–HAP40 complex was further purified by size-exclusion chromatography using a Superose 6 10/300 increase column (GE Healthcare) in running buffer (25 mM HEPES, 300 mM NaCl, 0.1% CHAPS and 1 mM DTT, pH 8.0). HTT–HAP40 eluted in one narrow-based peak and was concentrated with Amicon Ultra 100-kDa filters (Millipore).

HAP40 was purified from the HTT–HAP40 complex as follows. HTT–HAP40 bound to Strep beads was eluted with 25 mM HEPES, 300 mM NaCl, 0.05% *N*-dodecyl β-D-maltoside (DDM) and 2.5 mM desthiobiotin, pH 8.0. The eluate was concentrated using Amicon filters. To disrupt the HTT–HAP40 complex, DDM was added to a final concentration of 0.25%. After overnight incubation at 4 °C, the Strep eluate was further purified by size-exclusion chromatography using a Superose 6 10/300 increase column in running buffer (25 mM, 300 mM NaCl, 0.1% CHAPS and 1 mM DTT, pH 8.0) to separate HTT and HAP40. HAP40 eluted in one narrow-based peak and was concentrated with Amicon Ultra 30-kDa filters.

Rate-zonal ultracentrifugation. Sucrose gradients (5–20%) in 25 mM HEPES, 300 mM NaCl, 0.1% CHAPS, pH 8.0 were generated by an automatic gradient maker (Gradient Master, Biocomp Instruments). A volume of 120 μl of Flag-tag-purified HTT or Strep-tag-purified HTT–HAP40 complex was laid on top of the gradient and centrifuged at 39,000 r.p.m. for 16 h using a SW41 rotor in a Beckman ultracentrifuge. Fractions of the sucrose gradient were collected from the bottom of the tubes in fractions of 0.5 ml to be analysed by SDS–PAGE, Coomassie blue staining and western blotting.

Differential scanning fluorimetry. Protein thermostability was assessed by differential scanning fluorimetry[30]. Protein unfolding was monitored by the increase in the fluorescence of SYPRO Orange (Invitrogen). Before use, a 100 mM stock of the dye (stored at −20 °C) was diluted 1:20 in DMSO and directly added to the sample to a final concentration of 125 μM. The tested proteins were diluted in sample buffer (25 mM HEPES, 300 mM NaCl, 0.1% CHAPS, 1 mM DTT and 10% glycerol) to concentrations of 1.6 μM (HTT, HTT–HAP40 complex) and 2 μM (HAP40). The samples were heated up with a ramp rate of 1 °C min^{-1} over a temperature range of 15–95 °C using the qPCR System MX 3005 P (Stratagene). Measurements were performed in duplicate.

Transient expression of HAP40 and HAP40 fragments and interaction studies with HTT. Plasmids were generated expressing (under the control of the hCMV promoter) full-length HAP40, an N-terminal HAP40 fragment (HAP40-N, encoding residues 1–222), a C-terminal HAP40 fragment (HAP40-C, encoding residues 249–371) or a HAP40 fragment in which the central proline-rich region had been replaced by a flexible linker (HAP40Δ, encoding residues 1–222 linked by a $(\text{GGGGS})_3$ linker to residues 249–371). All HAP40 variants carried a C-terminal Twin-Strep tag.

B1.21 cells induced with doxycycline to express 17QHTT were transiently transfected with the plasmids using PEI transfection. At 48 h after transfection, the cells were collected by centrifugation and lysed in 25 mM HEPES, 300 mM NaCl, 0.5% Tween 20, 1× protease inhibitor, pH 8.0, followed by centrifugation (20,000*g*, 1 h). The supernatant was incubated with Magstrep beads (IBA) at 4 °C for 2 h and then washed three times with 25 mM HEPES, 300 mM NaCl, 0.02% Tween 20, pH 8.0. Thereafter, bound proteins were eluted using desthiobiotin in SDS loading buffer, followed by SDS–PAGE and western blot analysis using anti-Flag and anti-Strep antibodies for detection.

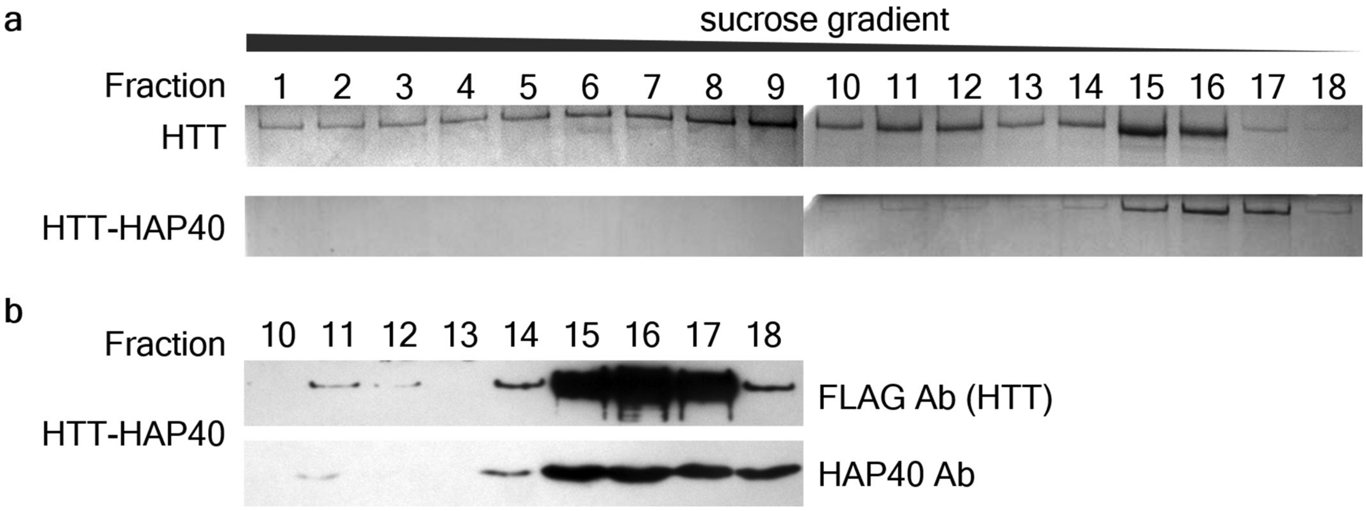
Cryo-EM sample preparation and data acquisition. Purified HTT–HAP40 complex was diluted to 0.5 mg ml^{-1} with 25 mM HEPES, 300 mM NaCl, 0.025% CHAPS, 1 mM DTT. A 4-μl volume of sample was applied to a Quantifoil gold grid suspended with monolayer graphene (Graphenea) and vitrified by plunge freezing in a liquid ethane–propane mixture using a Vitrobot Mark IV (FEI) with a blotting time of 5 s. Data collection was performed on a Titan Krios microscope (FEI) operated at 300 kV and equipped with a field emission gun, a Gatan GIF Quantum energy filter and a Gatan K2 Summit direct electron camera. The calibrated magnification was 105,000 in EFTEM mode, corresponding to a pixel size of 1.35 Å. Images were collected at a dose rate of 4 electrons $\text{Å}^{-2}\,\text{s}^{-1}$. Each exposure (8-s exposure time) comprised 16 sub-frames amounting to a total dose of 32 electrons $\text{Å}^{-2}\,\text{s}^{-1}$. Data was recorded using SerialEM[31] software and custom macros with defocus values ranging from −1.4 to −3 μm.

Image processing. Micrograph movie frame stacks were subjected to beam-induced motion correction by MotionCor2[32]. Most further processing was performed using RELION[33]. The contrast transfer function parameters for each micrograph were determined with CTFFIND4[34], and all micrographs with a resolution limit worse than 4 Å were discarded. Particles were initially picked with Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/Gautomatch_Brief_Manual.pdf) using a sphere as template, and extracted with a 160-pixel by 160-pixel box. Reference-free 2D-class averaging was performed reiteratively, keeping only particles with well-resolved 2D averages for initial model generation. To validate the *ab initio* model, 3D classification was performed using initial models generated by RELION, VIPER[35] or a simple sphere as reference. Identical 3D maps with detailed features were generated regardless of the reference used (Extended Data Fig. 8). 2D projections of this model were subsequently used as a reference to re-pick the particles. The resulting particles were subjected to reiterative reference-free 2D-class averaging. Strict selection of classes showing distinct structural features resulted in a particle subset used for further three-dimensional classification. The classes with identical detailed features were merged for further auto-refinement, applying a soft mask with six-pixel fall-off around the entire molecule, to produce the final density map with an overall resolution of 4 Å (Extended Data Fig. 2c). The resolution was estimated based on the gold-standard Fourier shell correlation (FSC) method using the 0.143 criterion[36]. The chirality of the final map was validated by model building of side chains within α-helices. All density maps were sharpened by applying a temperature factor that was estimated using post-processing in RELION. For visualization, the density maps were filtered based on the local resolution determined using half-reconstructions as input maps. Chimera[37] and PyMOL[38] were used for graphical visualizations.

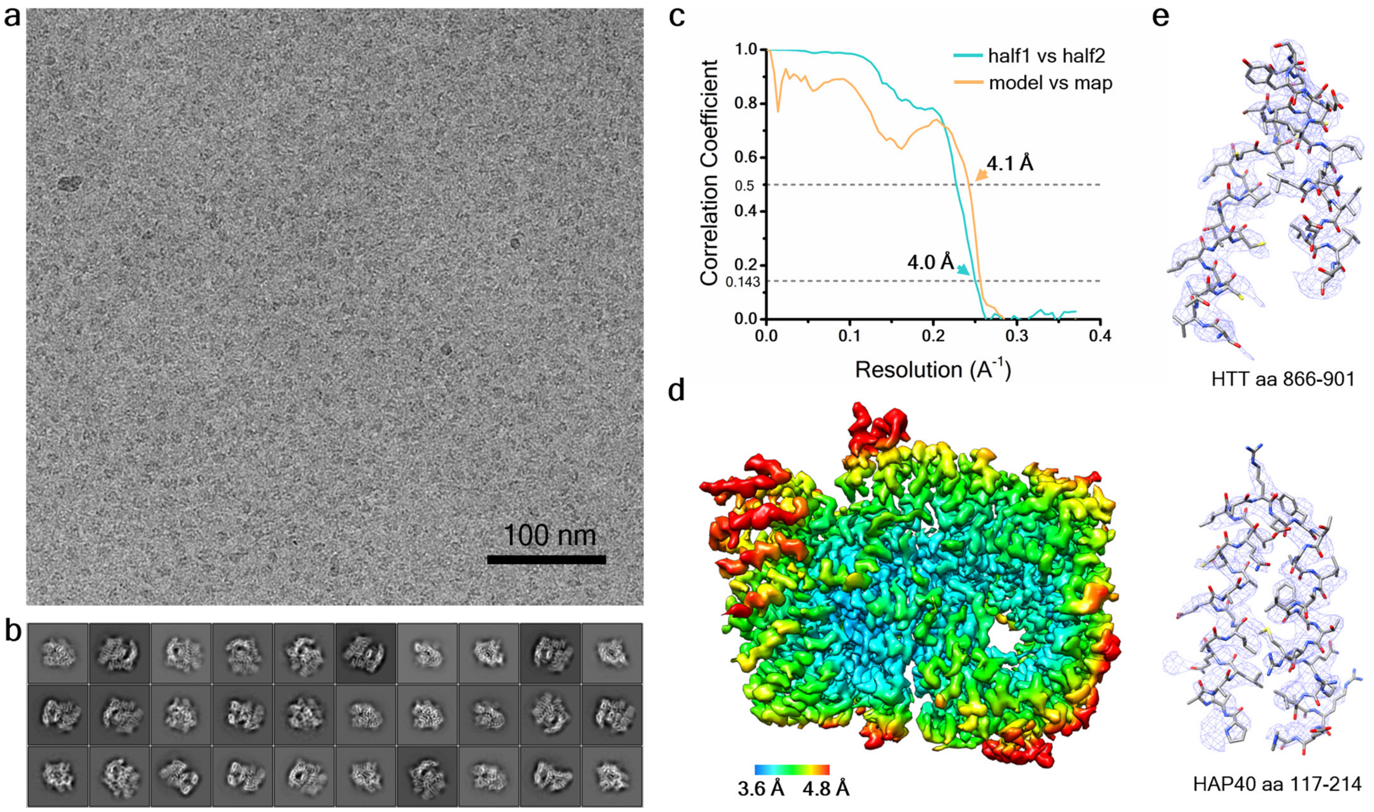
Model building. *Ab initio* modelling of the entire HTT–HAP40 complex was performed in Coot[39], using secondary structure predictions calculated by PSIPRED[19] and the densities of bulky side chains to determine registers of the residues. Some regions of HTT (1–90, 323–342, 403–660, 960–977, 1,049–1,057, 1,103–1,120, 1,158–1,222, 1,319–1,347, 1,372–1,418, 1,504–1,510, 1,549–1,556, 1,714–1,728, 1,855–1,881, 2,063–2,091, 2,325–2,347, 2,472–2,490, 2,580–2,582, 2,627–2,660, 2,681–2,687, 2,926–2,944 and 3,099–3,138) and HAP40 (1–41, 217–257, 300–313 and 365–371) were not built in the final model, as no well-resolved densities were present in the map. Map refinement was carried out using Phenix.real_space_refine[40] against the overall map at a resolution of 4 Å, with secondary structure and Ramachandran restraints. The final model was validated using MolProbity[41] (Extended Data Table 1).

Data availability. All data supporting the findings of this study are available within this paper. Source Data for Fig. 1 and Extended Data Fig. 2 and gel source images (Supplementary Fig. 1) are available with the online version of this paper. The cryo-EM map of the 17QHTT–HAP40 complex has been deposited in the Electron Microscopy Data Bank under accession code EMD-3984. The modelled structure of the 17QHTT–HAP40 complex has been deposited at the Protein Data Bank under accession code 6EZ8.

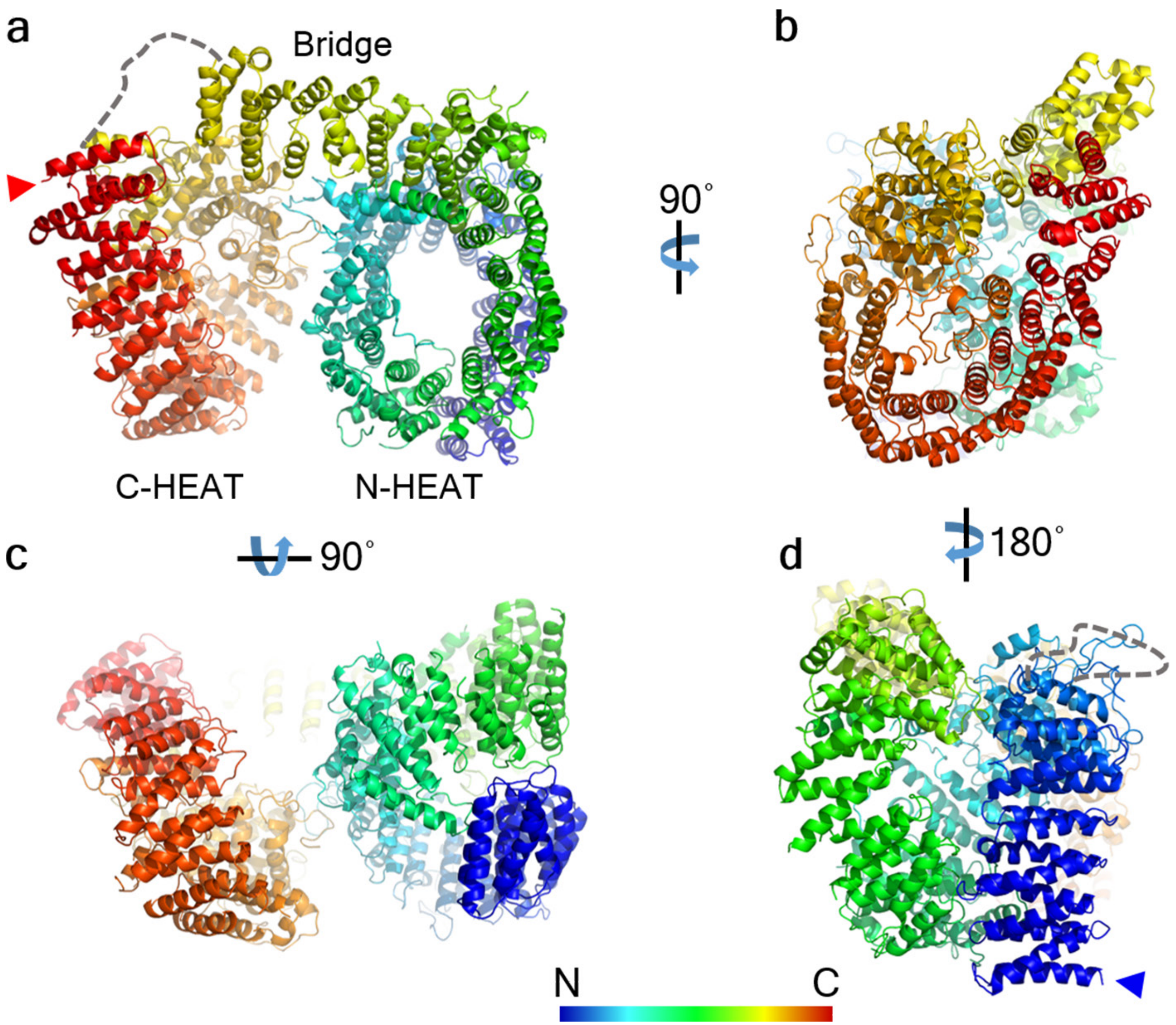
30. Vedadi, M. *et al.* Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl Acad. Sci. USA* **103,** 15835–15840 (2006).
31. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152,** 36–51 (2005).
32. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14,** 331–332 (2017).
33. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
34. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).
35. Moriya, T. *et al.* High-resolution single particle analysis from electron cryo-microscopy images using SPHIRE. *J. Vis. Exp.* **123,** e55448 (2017).
36. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9,** 853–854 (2012).
37. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
38. The PyMOL Molecular Graphics System v.1.8 (Schrödinger, 2015).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
40. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
41. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).



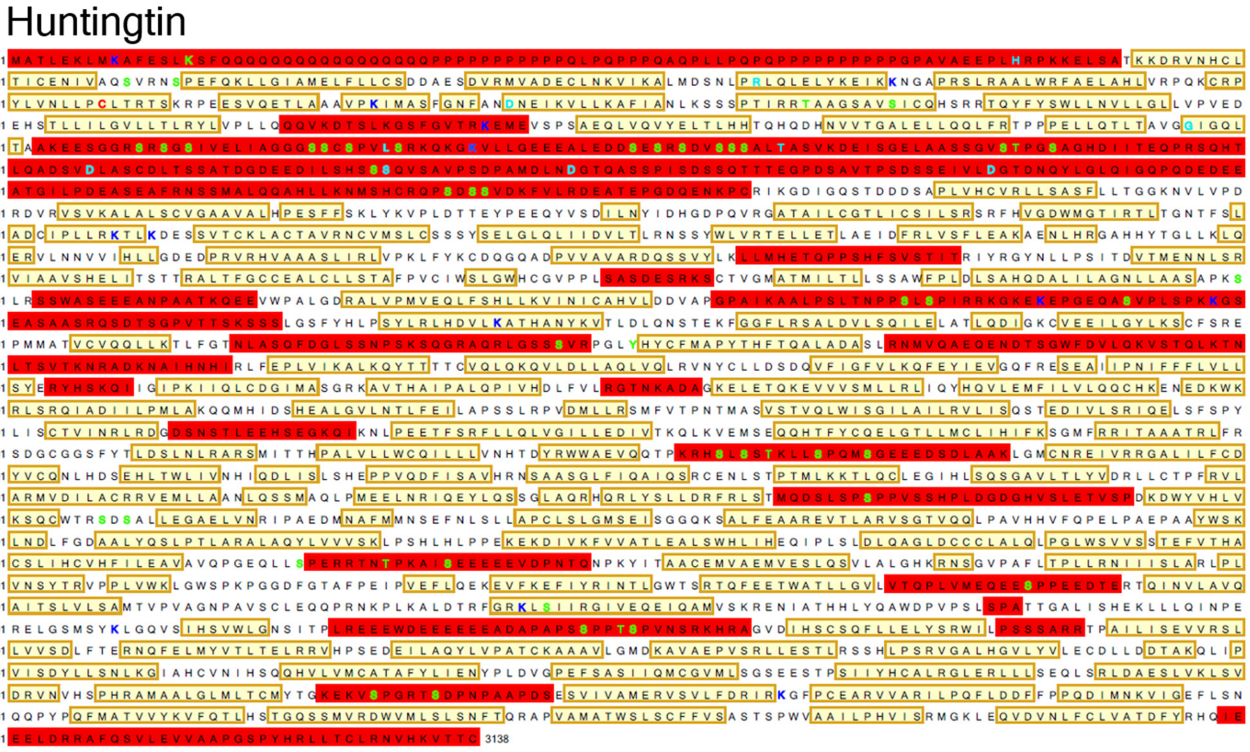
Extended Data Figure 1 | Sedimentation analysis by rate-zonal ultracentrifugation. a, Flag-tag purified HTT (top) and Strep-tag purified HTT–HAP40 complex (bottom) analysed by rate-zonal ultracentrifugation followed by SDS–PAGE and Coomassie staining. Twenty-five fractions from 5–20% sucrose gradients were collected from the bottom of the tube; fractions 1–18 are shown here. Whereas HTT alone was present in fractions 1–18, the HTT–HAP40 complex was found mainly in fractions 15–17, indicating lower conformational heterogeneity. **b**, Western blot analysis of fractions 10–18 of the HTT–HAP40 complex. Coomassie stainings and western blots are representative of three independent experiments with similar results. For gel source images, see Supplementary Fig. 1.



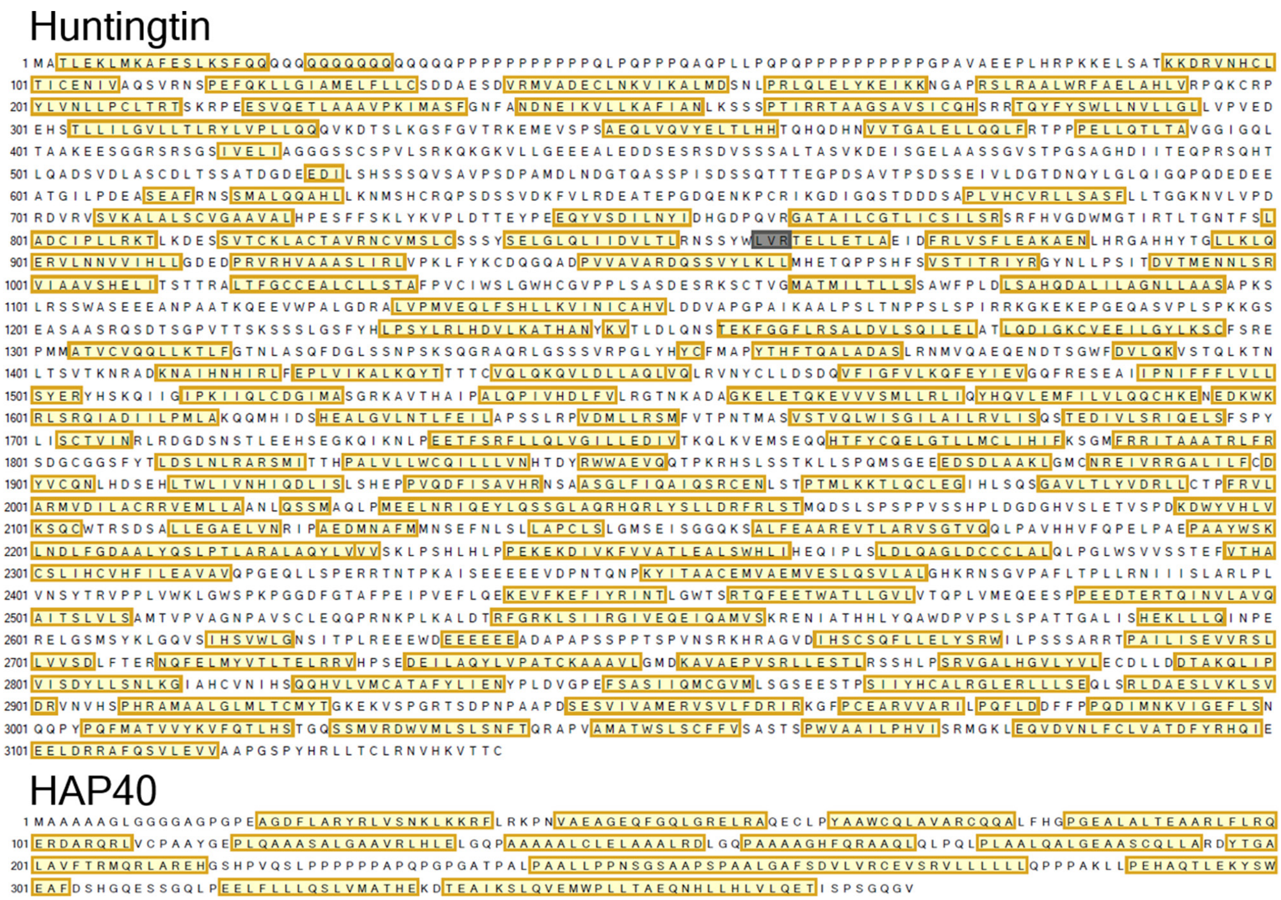
Extended Data Figure 2 | Cryo-EM analysis of the HTT–HAP40 complex. a, Representative micrograph of HTT–HAP40 complex. **b**, 2D class averages. **c**, FSC plots. Cyan, gold-standard FSC curve; orange, FSC curve calculated between the cryo-EM map and the refined atomic model. FSC cut-off values of 0.143 and 0.5 were used for the half versus half and model versus map comparisons, respectively. The initial and final numbers of micrographs and particles were 707 and 635 and 418,627 and 98,310, respectively. **d**, Final density map of the HTT–HAP40 complex, coloured according to local resolution. The map was low-pass filtered to 4 Å and sharpened with a *B*-factor of −174 Å^2. **e**, Detail of the electron density maps (mesh) for parts of HTT (top) and HAP40 (bottom). Source Data for the FSC plots are available online.



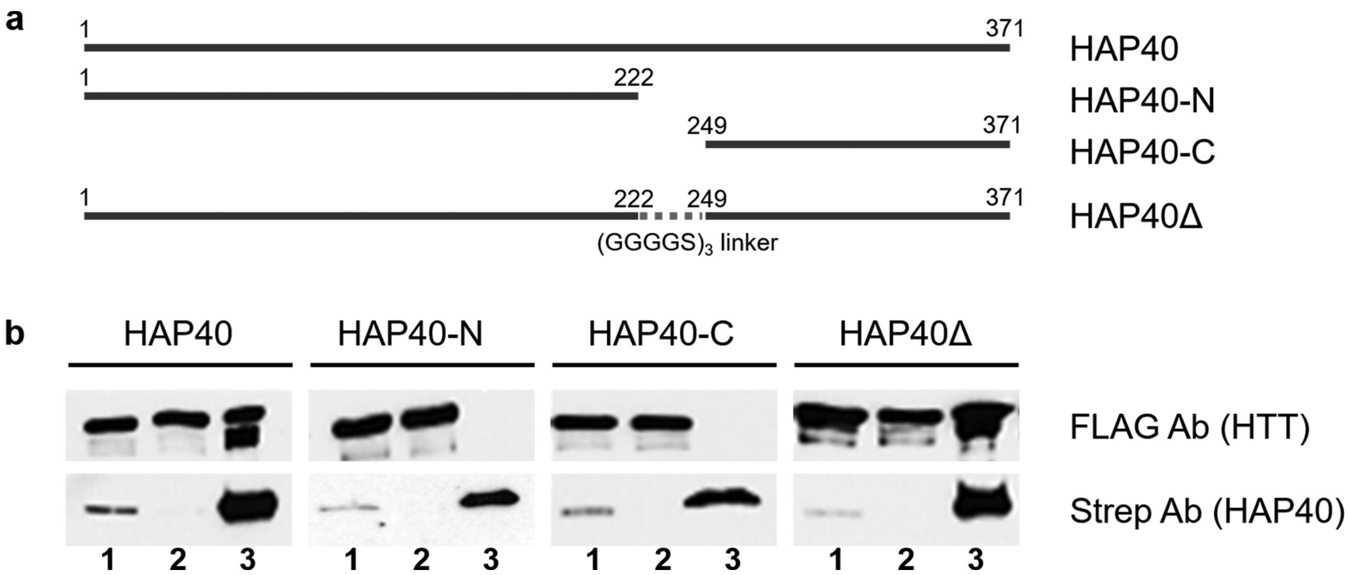
Extended Data Figure 3 | Atomic model of HTT within the HTT–HAP40 complex. a–d, The atomic model is shown in ribbon representation with a rainbow colour code from the N terminus (blue arrowhead in **d**) to the C terminus (red arrowhead in **a**). **a**–**d** show different views of the complex, as indicated. Dashed lines mark unresolved regions.



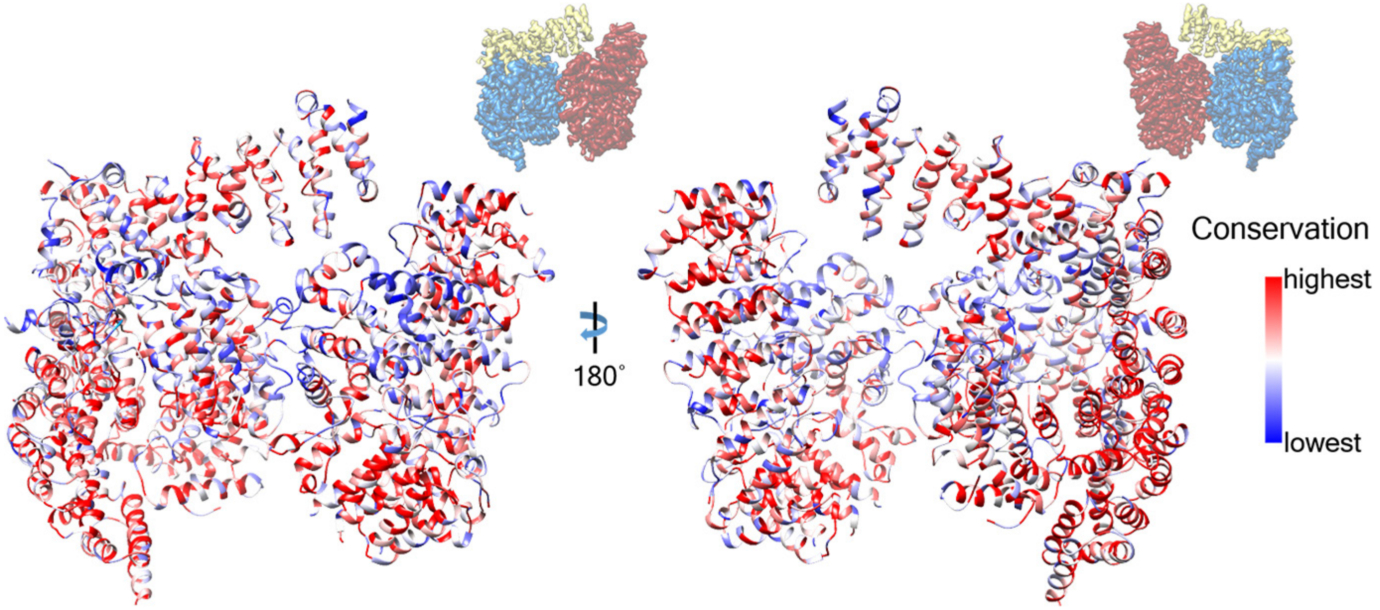
Extended Data Figure 4 | Amino acid sequences of 17QHTT and HAP40. Structural elements of the atomic models are indicated as follows: elements not visible in the model (red boxes), unstructured regions (no box) and α-helices (yellow boxes). The sites of previously reported protease cleavage and post-translational modifications of HTT[1,21,23,27] are indicated by text colour as follows: acetylation (dark blue), palmitoylation (red), phosphorylation (green) and proteolytic cleavage (cyan).



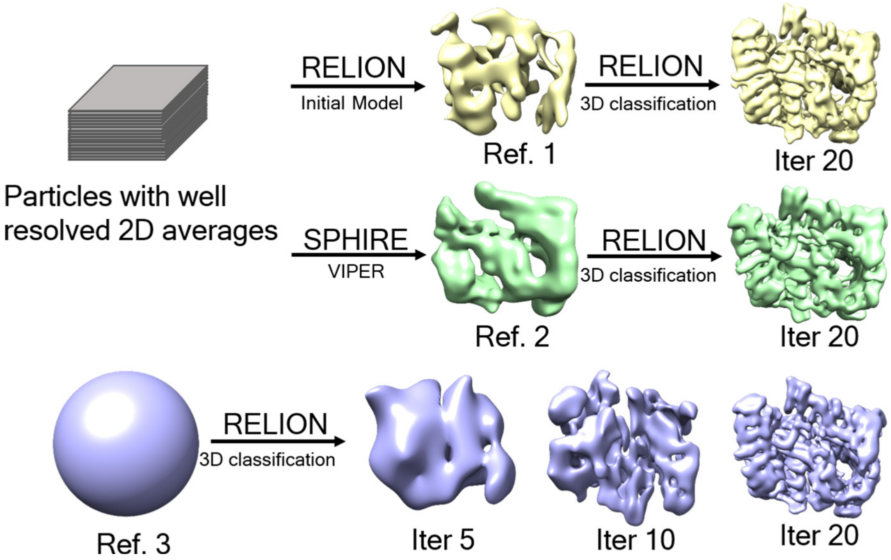
Extended Data Figure 5 | PSIPRED secondary-structure predictions for HTT and HAP40. Structural elements are indicated as follows: unstructured regions (no box), α-helices (yellow boxes) and β-sheet (grey box).



Extended Data Figure 6 | Truncation analysis of HAP40 binding to HTT. a, Schematic representation of the HAP40 constructs used in this study (all have C-terminal Strep-tags). **b**, HAP40 constructs were co-expressed with 17QHTT (bearing a C-terminal Flag-tag), immunoprecipitated using Strep-Tactin beads and analysed by western blot. Lanes are labelled as follows: 1, cell lysates; 2, cell lysates after incubation with Strep beads; 3, Strep-bead eluates. Note that full-length HAP40 and the construct lacking the central domain immunoprecipitate HTT, but constructs with deletion of the N- or C-terminal regions of HAP40 do not. Western blots are representative of two independent experiments with similar results. For gel source images, see Supplementary Fig. 1.

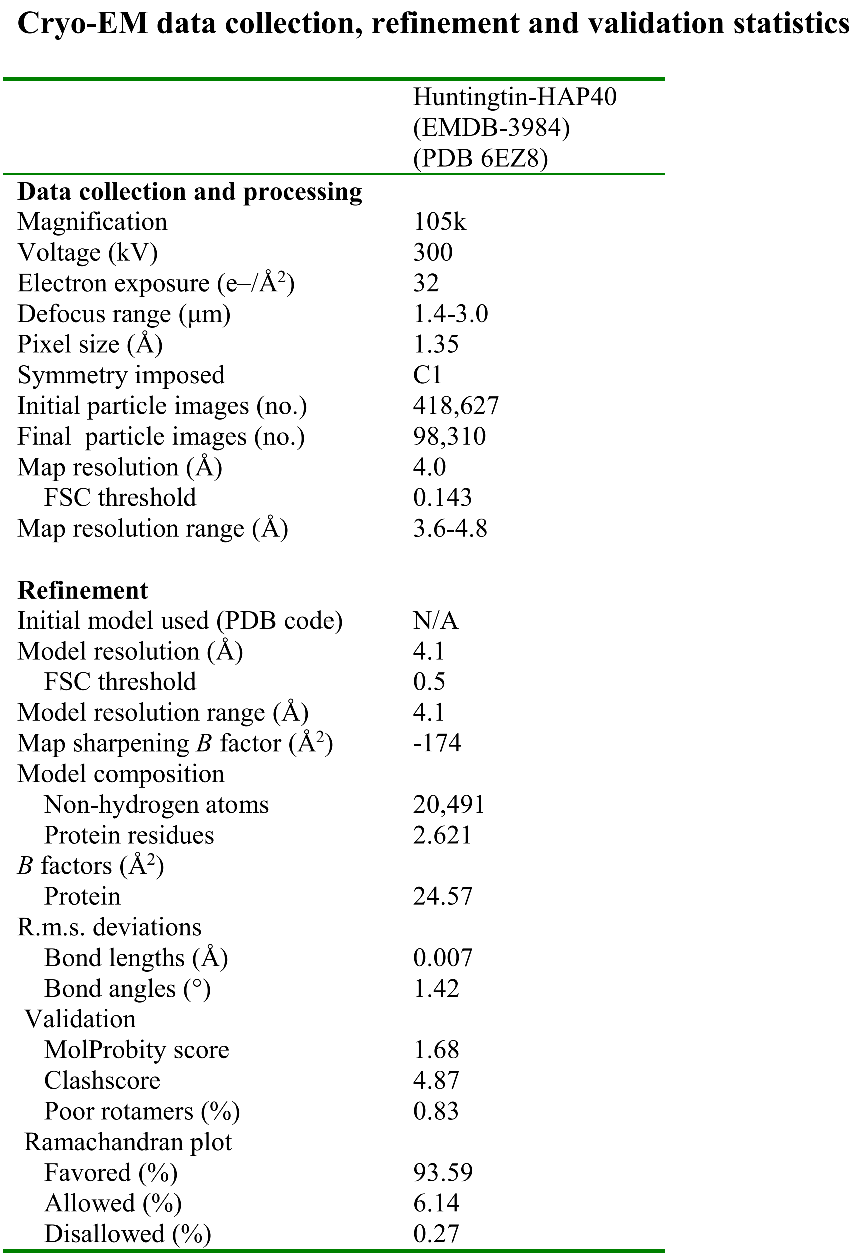


Extended Data Figure 7 | Evolutionary analysis of HTT. The human HTT model is shown in ribbon representation, coloured on the basis of sequence conservation across 16 metazoan species (*Homo sapiens, Rattus norvegicus, Mus musculus, Sus scrofa, Bos taurus, Canis lupus familiaris, Monodelphis domestica, Gallus gallus, Danio rerio, Tetraodon nigroviridis, Fugu rubripes, Ciona savignyi, Ciona intestinalis, Strongylocentrotus purpuratus, Tribolium castaneum, Apis mellifera*), using a previously reported sequence alignment[12]. The orientations of the HTT–HAP40 complex with respect to Fig. 2 are also indicated.



Extended Data Figure 8 | Workflow for initial model validation for 3D reconstruction of the HTT–HAP40 complex. A subset of particles with well-resolved 2D averages were used for initial model generation using RELION[33] or SPHIRE[35]. The resulting models were used as reference for 3D classification of all the good particles. A featureless sphere was also used as a classification reference. Most of the particles were classified to identical structures with sufficient detail, indicating no reference bias in the reconstruction.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics



Cryo-EM data collection, refinement and validation statistics

	Huntingtin-HAP40 (EMDB-3984) (PDB 6EZ8)
Data collection and processing	
Magnification	105k
Voltage (kV)	300
Electron exposure (e–/Å^2)	32
Defocus range (μm)	1.4-3.0
Pixel size (Å)	1.35
Symmetry imposed	C1
Initial particle images (no.)	418,627
Final particle images (no.)	98,310
Map resolution (Å)	4.0
FSC threshold	0.143
Map resolution range (Å)	3.6-4.8
Refinement	
Initial model used (PDB code)	N/A
Model resolution (Å)	4.1
FSC threshold	0.5
Model resolution range (Å)	4.1
Map sharpening *B* factor (Å^2)	-174
Model composition	
Non-hydrogen atoms	20,491
Protein residues	2.621
B factors (Å^2)	
Protein	24.57
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.42
Validation	
MolProbity score	1.68
Clashscore	4.87
Poor rotamers (%)	0.83
Ramachandran plot	
Favored (%)	93.59
Allowed (%)	6.14
Disallowed (%)	0.27

doi:10.1038/nature25773

Phase-plate cryo-EM structure of a biased agonist-bound human GLP-1 receptor–Gs complex

Yi-Lynn Liang[1]*, Maryam Khoshouei[2]*, Alisa Glukhova[1]*, Sebastian G. B. Furness[1], Peishen Zhao[1], Lachlan Clydesdale[1], Cassandra Koole[1], Tin T. Truong[1], David M. Thal[1], Saifei Lei[3,4], Mazdak Radjainia[1,5], Radostin Danev[2], Wolfgang Baumeister[2], Ming-Wei Wang[3,4,6], Laurence J. Miller[1,7], Arthur Christopoulos[1], Patrick M. Sexton[1,6] & Denise Wootten[1]

The class B glucagon-like peptide-1 (GLP-1) G protein-coupled receptor is a major target for the treatment of type 2 diabetes and obesity[1]. Endogenous and mimetic GLP-1 peptides exhibit biased agonism—a difference in functional selectivity—that may provide improved therapeutic outcomes[1]. Here we describe the structure of the human GLP-1 receptor in complex with the G protein-biased peptide exendin-P5 and a Gα_s heterotrimer, determined at a global resolution of 3.3 Å. At the extracellular surface, the organization of extracellular loop 3 and proximal transmembrane segments differs between our exendin-P5-bound structure and previous GLP-1-bound GLP-1 receptor structure[2]. At the intracellular face, there was a six-degree difference in the angle of the Gαs–α5 helix engagement between structures, which was propagated across the G protein heterotrimer. In addition, the structures differed in the rate and extent of conformational reorganization of the Gα_s protein. Our structure provides insights into the molecular basis of biased agonism.

The GLP-1R, a class B G protein-coupled receptor (GPCR), is a key incretin hormone receptor and an important target for the development of therapies for the treatment of type 2 diabetes and obesity[1]. Biased agonism is commonly observed at the GLP-1R[3–5], and exendin-P5 (ExP5) has been identified as a potent G protein-biased selective agonist of GLP-1R, with diminished coupling to β-arrestins[6] and a unique *in vivo* profile in animal models of diabetes[6]. The prevalence of GLP-1R biased agonism and its therapeutic implications make understanding of the phenomenon at molecular and structural levels crucial for the rational design of novel ligands.

Like all class B GPCRs, the GLP-1R contains a large extracellular N-terminal domain (NTD) and a seven-transmembrane helix bundle, with peptide binding spanning both domains; the NTD interaction positions the peptide N terminus within the receptor core to facilitate receptor activation[7]. Clinically used therapeutic agents, including exendin-4, contain an N-terminal sequence that is relatively conserved with that of the native peptide, GLP-1[8]. Notably, ExP5 shares a common C terminus with exendin-4, but possesses a unique N-terminal domain (Extended Data Fig. 1a) that interacts with the GLP-1R transmembrane core to promote receptor activation.

Cryo-electron microscopy (cryo-EM) has enabled researchers to determine the structures of GPCR complexes without the need to extensively modify the receptor[2,9]. A 4.1 Å full-length active structure of a wild-type rabbit GLP-1R was solved in complex with GLP-1 and heterotrimeric G$_s$ protein[2]. In addition, the full-length active structure of the calcitonin receptor (CTR) was solved to a similar global resolution in complex with a peptide agonist and G$_s$ protein[9] using phase contrast cryo-EM[10–12]. Here, we used Volta phase plate cryo-EM to determine the structure of an active state, human GLP-1R bound to ExP5 in complex with a heterotrimeric G$_s$ protein. The structure provides insights into the binding of ExP5 to the GLP-1R, with implications for receptor activation, G protein coupling and signalling for class B GPCRs.

To form an active, G protein-coupled complex, the GLP-1R was co-expressed with Gα_s, His-Gβ1, and Gγ2 in *Trichoplusia ni* (Tni) insect cells and stimulated with an excess of ExP5 in the presence of apyrase and the nanobody Nb35, which bridges the G protein α- and $\beta\gamma$-subunits. A dominant-negative Gα_s was used to enable the formation of a complex with improved stability. We characterized and purified the complex as described for the CTR[9] (Extended Data Figs 1b, 2a).

Following imaging and initial 2D classification (Extended Data Fig. 2b, c), 3D classification revealed that the majority of the complex had stable features. The exception was the Gα_s α-helical domain, the density of which was averaged out at higher resolution because it had substantial flexibility despite occupying a single predominant orientation (Fig. 1a). We used 184,672 particle projections to obtain a cryo-EM density map with nominal global resolution of 3.3 Å (Fig. 1a; Extended Data Fig. 2b).

An atomic resolution structure of the ExP5–GLP-1R–Gα_s heterotrimeric G protein complex was built into the map and refined to reveal global features similar to those observed in other class B GPCR structures[2,9,13–15]. Side chains of the majority of amino acid residues are clearly identifiable in the peptide, all of the transmembrane helices and the subunits of the G protein (Extended Data Fig. 3). Although linker region density between the NTD and the transmembrane core was visible in the cryo-EM map, it was less well-resolved than other receptor domains, suggesting substantial flexibility in the ExP5 bound state. Continuous density was observed for helix 8 (H8) and all intracellular and extracellular loops (ICLs and ECLs, respectively), with the exception of ICL3, which was not modelled. In addition, the cryo-EM map was poor for a region of four ECL3 residues (372–375) and therefore only the protein backbone was modelled in this region.

Within the NTD there was discontinuous density in the backbone for some regions. As such, the NTD structure bound to exendin(9–39)[16] was used to perform a rigid body fit into the density. N-terminal residues 24–30 and residues beyond E423 at the receptor C terminus were not resolved. The G protein was well resolved, allowing modelling of all G protein components (with the exception of the Gαs α-helical domain).

The extracellular NTD conformation differs between the three agonist-bound Gα_s heterotrimer class B GPCR structures (Extended Data Fig. 4a–c). Whereas multiple NTD conformations were evident for the CTR[9], a single predominant conformation was stabilized in both GLP-1R structures[2]. However, there were subtle differences in the relative positioning of the N terminus relative to the transmembrane

[1]Drug Discovery Biology and Department of Pharmacology, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville 3052, Victoria, Australia. [2]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany. [3]University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing 100049, China. [4]The National Center for Drug Screening and CAS Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. [5]FEI, 5651 GG Eindhoven, The Netherlands. [6]School of Pharmacy, Fudan University, Shanghai 201203, China. [7]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, Arizona 85259, USA.
*These authors contributed equally to this work.

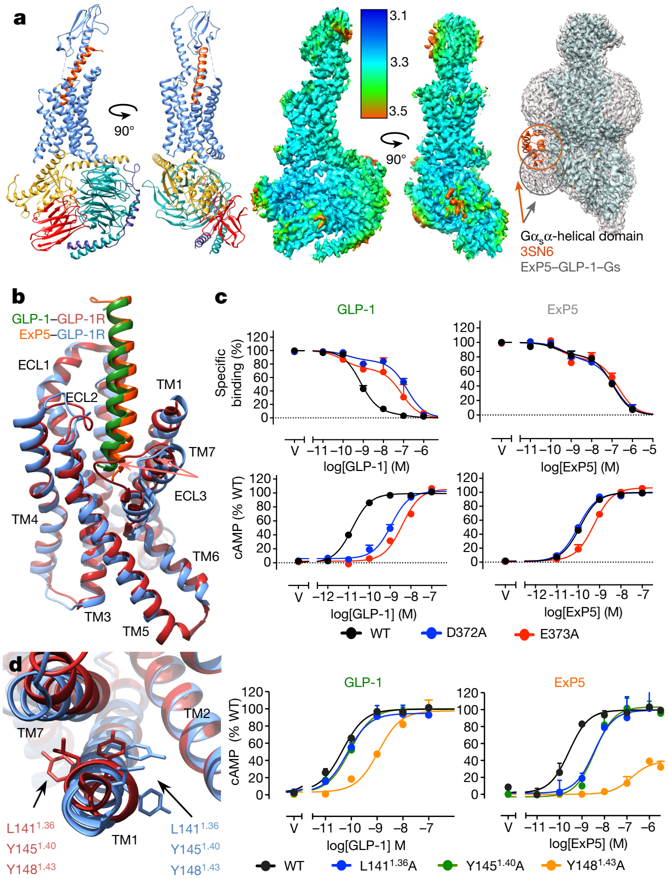


Figure 1 | The ExP5–GLP-1R–Gs cryo-EM structure reveals molecular details linked to GLP-1R biased agonism. a, Left, ExP5–GLP-1R–Gs structure after refinement in the cryo-EM map. Middle, cryo-EM density map coloured by local resolution (Å). Right, low-resolution cryo-EM map highlighting the predominant $G\alpha_s$ α-helical domain location in ExP5–GLP-1R–Gs (grey), compared to β2-AR–Gs (PDB:3SN6, orange). **b**, Transmembrane domain and peptide superimposition reveal backbone differences in ECL3, TM6, TM7 and TM1 when bound by GLP-1 relative to ExP5. ExP5 is located closer to TM1 than GLP-1. **c**, D372 and E373 in ECL3 are important for the pharmacology of GLP-1 and have a limited role in ExP5 affinity and signalling. WT, wild type; V, vehicle. **d**, Left, overlay of the GLP-1–GLP-1R deposited structure[2] (GLP-1R in red) and ExP5–GLP-1R (GLP-1R in blue) reveals a rotation in TM1 side chains. Right, L141$^{1.36}$, Y145$^{1.40}$ and Y148$^{1.43}$ mutations have a larger effect on ExP5-mediated than on GLP-1-mediated cAMP signalling. Whole-cell binding assays and cAMP signalling were assessed in CHOFlpIn cells and data are means + s.e.m. of four (for TM1) and six (for ECL3) independent experiments, performed in duplicate.

bundle that contribute to the positioning of the N termini of GLP-1 and ExP5 (Extended Data Fig. 4b). Notably, the 11-mer agonist-bound GLP-1R structure solved without the $G\alpha_s$ heterotrimer[15] displayed a unique NTD conformation relative to GLP-1 and ExP5 (Extended Data Fig. 4c). Collectively, these structures suggest that the binding of different peptide agonists alters the juxtaposition of the extracellular NTD and transmembrane bundle to regulate the ability of different peptides to activate class B GPCRs.

Compared to inactive class B GPCR transmembrane bundles, the GLP-1R in our structure undergoes similar macroscopic conformational transitions to those previously reported for the GLP-1-bound GLP-1R[2] and calcitonin-bound CTR[9] (Extended Data Fig. 4d–h). These include considerable movements in the extracellular ends of transmembrane (TM) helices 1, 6 and 7, required to open the bundle to accommodate peptide binding, and a large 15–16 Å movement of TM6 away from the central transmembrane domain axis that opens up the cytoplasmic face to accommodate G protein interaction (Extended Data Fig. 4d, f). These large conformational movements are coordinated around the highly conserved class B GPCR P$^{6.47}$XXG$^{6.50}$ motif in TM6, and G$^{7.50}$ in TM7 (Extended Data Fig. 4d). Nonetheless, there are notable differences in the extracellular face between the activated structures, particularly in the extent of movement of TM6, ECL3 and TM7, which probably reflect the distinct modes with which these ligands activate their respective receptors (Extended Data Fig. 4g, h).

ExP5 is a biased agonist relative to exendin-4[6]. Our pharmacological analysis revealed that ExP5 is also G protein-biased, with limited β-arrestin recruitment relative to GLP-1 (Extended Data Fig. 1d). Comparison of receptor occupancy with ligand potency and efficacy in cellular signalling assays showed that the bias of ExP5 arises primarily from enhanced efficacy in $G\alpha_s$-mediated cAMP signalling, rather than a loss of β-arrestin coupling (Extended Data Fig. 1e). Ligand binding and GTPγS studies performed in insect cells also support enhanced G protein efficacy of ExP5 relative to GLP-1 (Extended Data Fig. 1c). Thus, comparison of the GLP-1 and ExP5-bound GLP-1R–Gαsβγ structures provides insight into conformational differences that may be linked to biased agonism.

The largest distinctions between the GLP-1 and ExP5-bound GLP-1R transmembrane domains occur within TM1, the extracellular portions of TM6 and TM7, and the ECL3 conformation (Fig. 1b, Extended Data Fig. 5a), indicating that these domains may contribute to biased agonism. This is supported by earlier work identifying crucial roles for ECL3, and the extracellular helical boundaries of TM6 and TM7, within the GLP-1R for differential control of GLP-1R-mediated signalling[17]. Alanine scanning mutagenesis confirmed the importance of this domain for the differing pharmacological profiles of GLP-1 and ExP5 (Fig. 1c, Extended Data Table 1). Although some ECL3 residues (G377, R380) had similar roles in both GLP-1 and ExP5 function, the substitutions L379A, D372A and E373A substantially reduced GLP-1 affinity and signalling but had little effect on ExP5 function. Notably, the latter two residues lie within the region of ECL3 where the largest receptor backbone differences are observed between the two structures (Extended Data Fig. 5a), and alanine mutation converts the binding profile of GLP-1 to one that closely resembles the binding profile of ExP5 (Fig. 1c). Mutagenesis of these two residues also had a similar effect on the pharmacology of exendin-4, which has a bias profile similar to that of GLP-1 for these pathways (Extended Data Table 1). Moreover, mutation of L388$^{7.43}$ within the top of TM7 had a greater effect on GLP-1 signalling than on ExP5 signalling (Extended Data Fig. 5b), further supporting the importance of this region in biased agonism of GLP-1R.

There are additional differences between the ExP5-bound structure and the deposited GLP-1-bound GLP-1R structure, in the reported positioning of the TM1 kink and orientation of side chains in the extracellular half of TM1 (Extended Data Fig. 5c, Fig. 1d). The location of the TM1 kink in the 11-mer-bound GLP-1R and the agonist-bound CTR structures is equivalent to that observed in the ExP5-bound structure and an overlay of the ExP5-bound and GLP-1-bound GLP-1R cryo-EM maps reveals that they have similar backbone densities (Extended Data Fig. 5c). Although the limited density in the GLP-1 bound structure precludes confidence, the TM1 backbone can also be modelled in this common conformation, suggesting that the gross positioning of TM1 may be conserved, although comparison of the density maps indicates that the side chain positioning differs between the ExP5- and GLP-1-bound structures, possibly contributing to the biased agonism of ExP5. Indeed, in the deposited GLP-1-bound model, L141$^{1.36}$, Y145$^{1.40}$ and Y148$^{1.43}$ face towards TM7, whereas in the ExP5 structure they reside closer to TM2 (Fig. 1d). Mutation of these residues to alanine had a stronger effect on ExP5-mediated cAMP signalling than on GLP-1 signalling, supporting a role for TM1 in the control of signalling and an interaction between TM1 and TM7–ECL3–TM6 that manifests as altered $G\alpha_s$ efficacy and biased agonism between GLP-1 and ExP5.

Strong density was observed for the entirety of ExP5 extending from the NTD into deep within the transmembrane core (Extended Data Fig. 3).

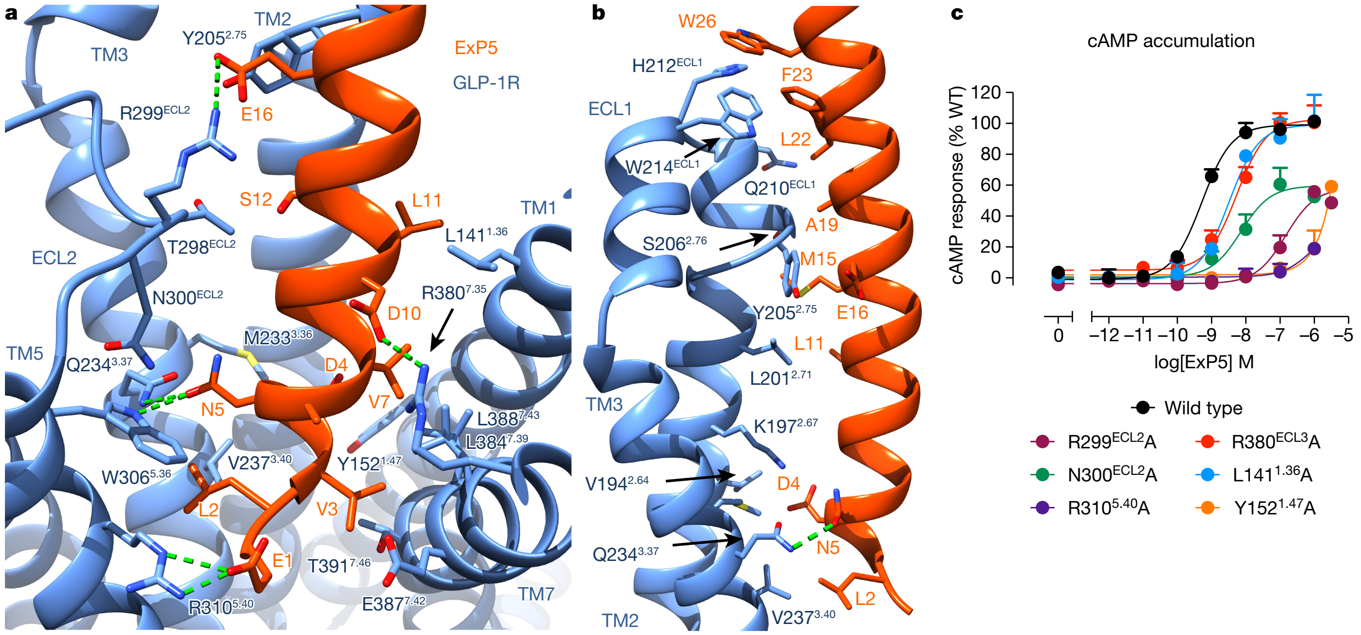


Figure 2 | The ExP5 binding site. a, Key interactions between ExP5 residues and TM1, TM3, TM5, TM7 and ECL2 of the GLP-1R transmembrane bundle (side chains located within 4 Å between the peptide (orange) and the GLP-1R (blue) are shown). ECL3 has been removed for clarity. **b**, Additional interactions formed by ExP5 with TM2, TM3 and ECL1. **c**, The functional effect on Gs-mediated cAMP accumulation following mutagenesis of key ExP5 residues that form interactions (highlighted in **a**) in the refined model supports the role of these residues in ExP5 interactions. cAMP signalling was assessed in CHOFlpIn cells and data are means + s.e.m. of four independent experiments performed in duplicate.

The peptide forms extensive interactions with residues in TMs 1, 2, 3, 5, 7 and all 3 ECLs (Fig. 2, Extended Data Table 2). Alanine mutagenesis confirmed the importance of key residues in the GLP-1R for ExP5 binding (Fig. 2c). Many of these residues lining the ExP5 binding site have previously been implicated as being important for binding of the cognate ligand, GLP-1[7,17–23].

E1 of ExP5 interacts with R310$^{5.40}$ of GLP-1R and is crucial for the ability of ExP5 to promote signalling through Gα_s, with R310$^{5.40}$A almost completely abolishing ExP5-mediated cAMP accumulation (Fig. 2a, c). Very clear density is evident for W306$^{5.36}$, which interacts directly with ExP5 through Van der Waal interactions with the aliphatic region of N5, as well as forming a direct hydrogen bond with N5 in the peptide. N5 also forms a hydrogen bond with Q234$^{3.37}$. N300^{ECL2} points down towards the receptor core within bonding distance of W306$^{5.36}$ and may participate in stabilizing these interactions. A series of contacts occur between residues in TM2 and ExP5, mainly through hydrophobic Van der Waals interactions with either hydrophobic residues or aliphatic regions of polar side chains (Fig. 2b, Extended Data Table 2). Peptide interactions also occur within ECL1, a region that has been implicated in peptide binding of other GLP-1R agonists[17,22] and ECL1 resides close to GLP-1 in the GLP-1-bound cryo-EM structure[2]. Van der Waals interactions are also formed between ExP5 and residues in TM1 and TM7 (Fig. 2a, Extended Data Table 2). In addition, two key electrostatic interactions are formed by R299^{ECL2} in ECL2 and R380$^{7.35}$ at the top of the TM7–ECL3 boundary with E16 and D10 of ExP5, respectively (Fig. 2a). These two residues also formed direct interactions with the 11-mer peptide agonist in the GLP-1R X-ray structure, interacting with a serine at position 8 (R299^{ECL2}) and an aspartic acid at position 9 (R380$^{7.35}$)[15]. D9 in the 11-mer is the equivalent of D10 in ExP5 and D15 in native GLP-1. An interaction between GLP-1 D15 and R380$^{7.35}$ has also been predicted by molecular dynamics simulations[17] and mutagenesis[23], but was not reported in the GLP-1-bound GLP-1R structure[2]. However, side chain densities were poorly resolved in this region of the deposited GLP-1–GLP-1R map; alternative modelling can preserve this interaction and therefore it is likely to be conserved across the three ligands for which structures are now available.

The GLP-1-bound GLP-1R cryo-EM structure also reported that R299^{ECL2} dips into the receptor core to form a direct interaction with H7 of GLP-1[2]. This modelling into the cryo-EM map is also ambiguous and contains an alternate positioning of W306$^{5.36}$ (required for R299^{ECL} to reach into the bundle) to the ExP5-bound and 11-mer-bound GLP-1R structures[15]. Because this positioning of W306$^{5.36}$ is not supported by density, and the described interaction of R299^{ECL2} is highly energetically unfavourable, we hypothesize that W306$^{5.36}$ is more likely to reside in a similar orientation to that observed in the ExP5- and 11-mer-bound structures, supported by good density in these maps. This orientation would promote interactions of R299^{ECL2} with GLP-1 higher up in the peptide.

Owing to the limited density available to define GLP-1 interactions in the GLP-1-bound GLP-1R cryo-EM map, it is difficult to assess direct differences in peptide interactions between the GLP-1- and ExP5-bound structures by relying on the structures alone. Nonetheless, alanine mutation of residues lining the ExP5-binding pocket (highlighted in Fig. 2c, Extended Data Table 1) confirmed a likely overlap of GLP-1R residues involved in interactions with GLP-1 and ExP5, with previous publications highlighting the importance of Y205$^{2.75}$, R299^{ECL2}, N300 ^{ECL2}, R380$^{7.35}$ and R310$^{5.40}$ in GLP-1 affinity and signalling[1,17,20,21], and our results confirming their importance for ExP5 binding (Fig. 2). The nature of these interactions is likely to differ, owing to the variations in peptide sequence and consequent receptor interactions, as highlighted by the TM1, TM7 and ECL3 mutagenesis in this study.

Class B GPCRs contain a number of highly conserved transmembrane polar residues that participate in key hydrogen bond interactions for receptor integrity and maintenance of the apo state. A central polar network formed by residues R$^{2.60}$, N$^{3.43}$, H$^{6.52}$ and Q$^{7.49}$ is located just below the peptide binding site in the ExP5-bound structure[24,25] (Extended Data Fig. 6). Two highly conserved class B GPCR polar networks (TM2–TM3–TM6–TM7: H$^{2.50}$, E$^{3.50}$, T$^{6.42}$, Y$^{7.57}$ and TM2–TM6–TM7–H8: R$^{2.46}$, R/K$^{6.37}$, N$^{7.61}$, E$^{8.41}$) at the cytoplasmic face lock the base of the receptor in an inactive conformation[21,25]. Located between the central hydrogen bond network and the TM2–TM3–TM6–TM7 network is a cluster of conserved residues that form

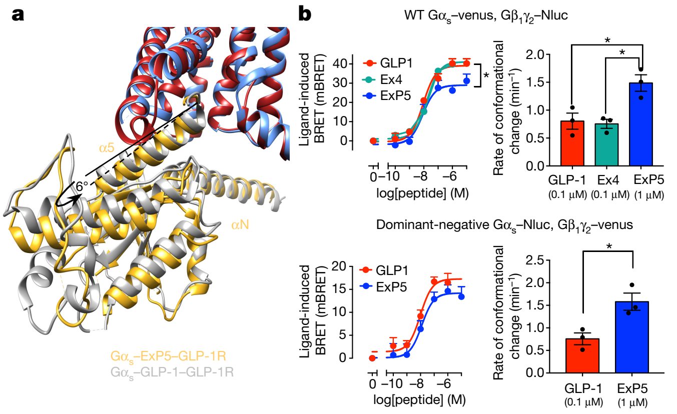


Figure 3 | Comparison of GLP-1R-mediated G protein conformation in GLP-1-bound and ExP5-bound receptors. a, Superimposition of the GLP-1R bundle bound by GLP-1 and by ExP5 reveals distinct angles of $G\alpha_s$ α5 engagement (6° measured using pisco). **b**, Top, BRET measurements show distinct conformational rearrangements between the $G\alpha_s$ α-helical domain and Gγ when the GLP-1R is activated by ExP5, relative to activation by GLP-1 or exendin-4. This is associated with a faster rate of rearrangement at equi-occupant ligand concentrations. Bottom, similar differences are observed with the dominant-negative $G\alpha_s$. Data are means + s.e.m. (left panels) or mean ± s.e.m. (right panels) of three independent experiments performed in triplicate. * $P < 0.05$ by one-way analysis of variance followed by Tukey's multiple comparisons post-test.

hydrophobic packing interactions in the inactive state, stabilizing the TM6 $P^{6.47}XXG^{6.50}$ motif in an inactive conformation (Extended Data Fig. 6). Upon peptide binding, a reorganization of the GLP-1R central hydrogen bond network is associated with destabilization within TM6 around the $P^{6.47}XXG^{6.50}$ motif and a major rearrangement of the central hydrophobic network to form a new packing arrangement that stabilizes the active state (Extended Data Fig. 6, Supplementary Video 1). These major rearrangements break two hydrogen bond networks at the bottom of the receptor, facilitating movement of TM6 away from the transmembrane bundle to create a cavity for G protein binding (Extended Data Figs 6, 7b–d, Supplementary Video 1). $Y^{7.57}$ and $H^{2.50}$ are released from their ground state constraints and reorganize to form part of the hydrophobic network that stabilizes the active state. $E^{3.50}$ maintains a hydrogen bond interaction with $H^{2.50}$, further stabilizing this active conformation.

The GLP-1R active conformation is stabilized by extensive interactions with the $G\alpha_s$ heterotrimeric protein (Extended Data Fig. 7). The receptor–$G\alpha_s$ heterotrimer interface is formed by residues located in TM2, TM3, TM5, TM6, TM7, ICL1, ICL2, ICL3 and H8 of the GLP-1R, and the α5 and αN regions of $G\alpha_s$ and Gβ (Extended Data Table 3).

H8 in all active structures is amphipathic, with bulky aromatic residues on the membrane-proximal face heavily buried in the detergent micelle. Direct interactions of H8 and ICL1 with Gβ are conserved across class B GPCR G protein structures[2,9] (Extended Data Fig. 7e) and these are summarized in Extended Data Table 3. Though the importance of these interactions for the GLP-1R is unclear, truncation of H8 in the CTR reduced receptor expression and peptide-mediated cAMP efficacy, suggesting that receptor–Gβ interactions are important for class B GPCR function[9].

In all structures, the most extensive G protein contacts are formed by the α5 helix of the $G\alpha_s$ Ras-like domain, which inserts into the central GLP-1R transmembrane bundle cytoplasmic cavity formed by the 15 Å outward movement of TM6 (Extended Data Fig. 7). These contacts consist of both polar and hydrophobic Van der Waals interactions and there is, generally, a common interaction pattern between $G\alpha_s$ and the available active class B GPCRs (Extended Data Table 3).

Superimposition of the G proteins of the GLP-1- and ExP5-bound GLP-1R structures reveals only relatively small differences in the receptor-complexed $G\alpha_s$Ras and Gβγ domains (Extended Data Fig. 7f). The largest change was a 4 Å variance in the conformation of the $G\alpha_s$αN domain at its N terminus, which may reflect a ligand-dependent difference in conformation.

Superimposition of the transmembrane domains of the GLP-1R in the GLP-1- and ExP5- bound structures reveals that, although there are limited differences in the overall $G\alpha_s$Ras and Gβγ conformations, there is a six-degree variance in the angle at which the $G\alpha_s$ α5 helix engages in the GLP-1R cytoplasmic cavity. This results in an overall rotation of the G protein in the ExP5-bound structure relative to the GLP-1-bound structure (Fig. 3a, Extended Data Fig. 7f). Notably, when ExP5 is bound to the GLP-1R, the α4 helix and β3 strand are located further from the receptor core, and no interactions are observed between the α4 helix and the GLP-1R intracellular face, whereas there are potential contacts for the GLP-1-bound structure[2]. In addition, the αN–β3 loop of Gαs is located further from ICL2 of the GLP-1R in the ExP5-bound structure; although these side chains are still within bonding distance, their interactions are likely to be weaker than those induced by GLP-1 binding. Notably, there was only very limited density within the backbone for residues in the bottom of TM5–ICL3 (residues 337–343) in the ExP5-bound structure, such that this region is not visible in high-resolution maps, whereas this backbone density was clearly visible for the GLP-1-bound structure (Extended Data Fig. 5d). This suggests that ICL3 of the GLP-1R is less flexible in the GLP-1- and G protein-bound state than in the ExP5- and G protein-bound state.

There are multiple lines of evidence that differences in ligand–receptor conformation propagate to G protein conformation[26,27]. Direct assessment of conformational rearrangement between $G\alpha_s$ and Gγ, using a bioluminescence resonance energy transfer (BRET) assay, revealed that ExP5 promotes a faster conformational change within $G\alpha_s$ than do GLP-1 or exendin-4 at equi-occupant concentrations, accompanied by a lower BRET maximal signal (E_{max}) at saturating concentrations of peptide (Fig. 3b). Together with the structural data, these results are consistent with the distinct flexibilities of the bottom of TM5 and within ICL3 altering the conformational positioning of the $G\alpha_s$ α-helical domain and increasing the rate of G protein activation. Collectively, this may contribute to the enhanced $G\alpha_s$ protein-mediated efficacy of ExP5 that is a key element of its biased agonism.

In conclusion, the structure of the ExP5–GLP-1R–$G\alpha_s$ complex provides insights into the structural reorganization of class B GPCRs upon peptide activation, as well as the distinct engagement of GLP-1R agonists with differential signalling bias. Our results highlight that even when ligands share a common G protein transducer, differences in the mode of G protein binding can have consequences for conformational changes in the G protein that are linked to activation. The findings increase our understanding of biased agonism and may contribute to the rational design of novel therapeutics that target the GLP-1R.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 September 2017; accepted 17 January 2018.

Published online 21 February 2018.

1. de Graaf, C. *et al.* Glucagon-like peptide-1 and its class B G protein-coupled receptors: a long march to therapeutic successes. *Pharmacol. Rev.* **68,** 954–1013 (2016).
2. Zhang, Y. *et al.* Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **546,** 248–253 (2017).
3. Hager, M. V., Clydesdale, L., Gellman, S. H., Sexton, P. M. & Wootten, D. Characterization of signal bias at the GLP-1 receptor induced by backbone modification of GLP-1. *Biochem. Pharmacol.* **136,** 99–108 (2017).
4. Koole, C. *et al.* Allosteric ligands of the glucagon-like peptide 1 receptor (GLP-1R) differentially modulate endogenous and exogenous peptide responses in a pathway-selective manner: implications for drug screening. *Mol. Pharmacol.* **78,** 456–465 (2010).
5. Wootten, D. *et al.* Differential activation and modulation of the glucagon-like peptide-1 receptor by small molecule ligands. *Mol. Pharmacol.* **83,** 822–834 (2013).
6. Zhang, H. *et al.* Autocrine selection of a GLP-1R G-protein biased agonist with potent antidiabetic effects. *Nat. Commun.* **6,** 8918 (2015).

7. Mann, R. *et al.* Peptide binding at the GLP-1 receptor. *Biochem. Soc. Trans.* **35,** 713–716 (2007).
8. Manandhar, B. & Ahn, J. M. Glucagon-like peptide-1 (GLP-1) analogs: recent advances, new possibilities, and therapeutic implications. *J. Med. Chem.* **58,** 1020–1037 (2015).
9. Liang, Y. L. *et al.* Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature* **546,** 118–123 (2017).
10. Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. M. & Baumeister, W. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proc. Natl Acad. Sci. USA* **111,** 15635–15640 (2014).
11. Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun.* **8,** 16099 (2017).
12. Khoshouei, M. *et al.* Volta phase plate cryo-EM of the small protein complex Prx3. *Nat. Commun.* **7,** 10534 (2016).
13. Siu, F. Y. *et al.* Structure of the human glucagon class B G-protein-coupled receptor. *Nature* **499,** 444–449 (2013).
14. Hollenstein, K. *et al.* Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **499,** 438–443 (2013).
15. Jazayeri, A. *et al.* Crystal structure of the GLP-1 receptor bound to a peptide agonist. *Nature* **546,** 254–258 (2017).
16. Runge, S., Thøgersen, H., Madsen, K., Lau, J. & Rudolph, R. Crystal structure of the ligand-bound glucagon-like peptide-1 receptor extracellular domain. *J. Biol. Chem.* **283,** 11340–11347 (2008).
17. Wootten, D. *et al.* The extracellular surface of the GLP-1 receptor is a molecular trigger for biased agonism. *Cell* **165,** 1632–1643 (2016).
18. Coopman, K. *et al.* Residues within the transmembrane domain of the glucagon-like peptide-1 receptor involved in ligand binding and receptor activation: modelling the ligand-bound receptor. *Mol. Endocrinol.* **25,** 1804–1818 (2011).
19. Dods, R. L. & Donnelly, D. The peptide agonist-binding site of the glucagon-like peptide-1 (GLP-1) receptor based on site-directed mutagenesis and knowledge-based modelling. *Biosci. Rep.* **36,** e00285 (2015).
20. Koole, C. *et al.* Second extracellular loop of human glucagon-like peptide-1 receptor (GLP-1R) has a critical role in GLP-1 peptide binding and receptor activation. *J. Biol. Chem.* **287,** 3642–3658 (2012).
21. Wootten, D. *et al.* Key interactions by conserved polar amino acids located at the transmembrane helical boundaries in class B GPCRs modulate activation, effector specificity and biased signalling in the glucagon-like peptide-1 receptor. *Biochem. Pharmacol.* **118,** 68–87 (2016).
22. Yang, D. *et al.* Structural determinants of binding the seven-transmembrane domain of the glucagon-like peptide-1 receptor (GLP-1R). *J. Biol. Chem*. **291,** 12991–13004 (2016).
23. Moon, M. J. *et al.* Ligand binding pocket formed by evolutionarily conserved residues in the glucagon-like peptide-1 (GLP-1) receptor core domain. *J. Biol. Chem.* **290,** 5696–5706 (2015).
24. Wootten, D. *et al.* A hydrogen-bonded polar network in the core of the glucagon-like peptide-1 receptor is a fulcrum for biased agonism: lessons from class B crystal structures. *Mol. Pharmacol.* **89,** 335–347 (2016).
25. Wootten, D., Simms, J., Miller, L. J., Christopoulos, A. & Sexton, P. M. Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family B G protein-coupled receptor conformations. *Proc. Natl Acad. Sci. USA* **110,** 5211–5216 (2013).
26. Furness, S. G. B. *et al.* Ligand-dependent modulation of G protein conformation alters drug efficacy. *Cell* **167,** 739–749.e11 (2016).
27. Gregorio, G. G. *et al.* Single-molecule analysis of ligand efficacy in β2AR-G-protein activation. *Nature* **547,** 68–73 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements The work was supported by the Monash University Ramaciotti Centre for Cryo-Electron Microscopy, the National Health and Medical Research Council of Australia (NHMRC) project grants (1061044, 1065410, 1120919 and 1126857), NHMRC program grant (1055134), Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12020347) and Shanghai Science and Technology Development Fund (15DZ2291600). P.M.S., A.C., D.W. and C.K. are NHMRC Principal Research, Senior Principal Research, Career Development and CJ Martin Fellows, respectively. S.L. received the Postgraduate Overseas Study Fellowship from CAS. We thank J. Plitzko, G. Christopoulos, V. Julita, J. Michaelis, X. Zhang, P. Thompson and M. Liu for assay and technical support and B. Kobilka for technical advice and comments on the manuscript.

Author Contributions Y.-L.L. established the GLP-1R complex expression and purification strategy, expressed and purified the complex, and performed negative stain EM and data acquisition/analysis; Y.-L.L. and M.R. performed preliminary cryo-EM screening; M.K. performed cryo-sample preparation and phase plate imaging to acquire EM data and performed EM map calculations; A.G. built the model and performed refinement; A.G., C.K. and D.M.T. performed pharmacological assays; L.C., T.T.T. and S.L. performed the mutagenesis studies; S.G.B.F. and P.Z. designed and performed the G protein BRET assays; R.D. and W.B. organized and developed the Volta phase plate cryo-EM data acquisition strategy; Y.-L.L., M.K., A.G., S.G.B.F., P.Z., L.C., C.K., D.M.T., T.T.T., S.L., A.C., P.M.S. and D.W. performed data analysis; S.G.B.F., P.Z., C.K., A.C., L.J.M., M.-W.W. and A.C. assisted with data interpretation and preparation of the manuscript; Y.L.L., M.K., A.G., P.M.S. and D.W. interpreted data and wrote the manuscript; P.M.S. and D.W. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.W. (denise.wootten@monash.edu) and P.M.S. (Patrick.sexton@monash.edu).

Reviewer Information *Nature* thanks R. Glaeser, F. Marshall and J. Mayer for their contribution to the peer review of this work.

METHODS

Constructs. The human GLP-1R was unmodified with the exception of replacing the native signal peptide with that of haemagglutinin (HA) to enhance receptor expression and the addition of affinity tags (an N-terminal Flag tag epitope and a C-terminal 8× His tag); both tags are removable by 3C protease cleavage. The construct was generated in both mammalian and insect cell expression vectors. These modifications did not alter receptor pharmacology (Extended Data Fig. 1b). A dominant-negative $G\alpha_s$ ($DNG\alpha_s$) construct was generated by site-directed mutagenesis to incorporate mutations that alter nucleotide handling (S54N[28] and G226A[29]), stabilize the G_0 state (E268A[30]) and substitute residues from $G\alpha_{i2}$ (N271K, K274D, R280K, T284D and I285T[31,32]) that are reported to improve the dominant-negative effect, presumably by stabilizing interactions with the $\beta\gamma$ subunits.

Insect cell expression. Human GLP-1R, human $DNG\alpha_s$, and His_6-tagged human $G\beta_1$ and $G\gamma_2$ were expressed in Tni insect cells (Expression systems) using baculovirus. Cell cultures were grown in ESF 921 serum-free medium (Expression Systems) to a density of 4 million cells per ml and then infected with three separate baculoviruses at a ratio of 2:2:1 for GLP-1R, $DNG\alpha_s$ and $G\beta_1\gamma_2$. The culture was collected by centrifugation 60 h after infection and cell pellets were stored at −80 °C.

Complex purification. Cell pellets were thawed in 20 mM HEPES pH 7.4, 50 mM NaCl, 2 mM $MgCl_2$ supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche). Complex formation was initiated by addition of 1 μM ExP5 (China Peptides), Nb35–His (10 μg/ml) and apyrase (25 mU/ml, NEB); the suspension was incubated for 1 h at room temperature. Membranes were collected by centrifugation at 30,000*g* for 30 min, and complex was solubilized from membrane using 0.5% (w/v) lauryl maltose neopentyl glycol (LMNG, Anatrace) supplemented with 0.03% (w/v) cholesteryl hemisuccinate (CHS, Anatrace) for 2 h at 4 °C in the presence of 1 μM ExP5 and apyrase (25 mU/ml, NEB). Insoluble material was removed by centrifugation at 30,000*g* for 30 min and the solubilized complex was immobilized by batch binding to M1 anti-Flag affinity resin in the presence of 3 mM $CaCl_2$. The resin was packed into a glass column and washed with 20 column volumes of 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 1 μM ExP5, 0.01% (w/v) LMNG and 0.006% (w/v) CHS before bound material was eluted in buffer containing 5 mM EGTA and 0.1 mg/ml FLAG peptide. The complex was then concentrated using an Amicon Ultra Centrifugal Filter (MWCO, 100 kDa) and subjected to size-exclusion chromatography on a Superdex 200 Increase 10/300 column (GE Healthcare) that was pre-equilibrated with 20 mM HEPES pH 7.4, 100 mM NaCl, 2 mM $MgCl_2$, 1 μM ExP5, 0.01% (w/v) MNG and 0.006% (w/v) CHS to separate complex from contaminants. Eluted fractions consisting of receptor and G-protein complex were pooled and concentrated. The final yield of purified complex was approximately 0.2 mg per litre of insect cell culture.

SDS–PAGE and western blot analysis. Samples collected from size-exclusion chromatography were analysed by SDS–PAGE and western blot. For SDS–PAGE, precast gradient TGX gels (Bio-Rad) were used. Gels were either stained by Instant Blue (Expedeon) or immediately transferred to PVDF membrane (Bio-Rad) at 100 V for 1 h. The proteins on the PVDF membrane were probed with two primary antibodies, rabbit anti-$G\alpha_s$ C-18 antibody (cat. no. sc-383, Santa Cruz) against the $G\alpha_s$ subunit and mouse penta-His antibody (cat. no. 34660, QIAGEN) against His tags. The membrane was washed and incubated with secondary antibodies (680RD goat anti-mouse and 800CW goat anti-rabbit, LI-COR). Bands were imaged using an infrared imaging system (LI-COR Odyssey Imaging System).

Preparation of vitrified specimen. EM grids (Quantifoil, 200 mesh copper R1.2/1.3) were glow discharged for 30 s in high pressure air using Harrick plasma cleaner. Four microlitres of sample at 1.3 mg/ml was applied to the grid in the Vitrobot chamber (FEI Vitrobot Mark IV). The Vitrobot chamber was set to 100% humidity at 4 °C. The sample was blotted for 5 s with a blot force of 20 and then plunged into propane–ethane mixture (37% ethane and 63% propane).

Data acquisition. Data were collected on a Titan Krios microscope operated at 300 kV (Thermo Fisher Scientific equipped with a Gatan Quantum energy filter, a Gatan K2 summit direct electron detector (Gatan) and a Volta phase plate (Thermo Fisher Scientific)). Videos were recorded in EFTEM nanoprobe mode, with 50-μm C2 aperture, at a calibrated magnification of 47,170 corresponding to a magnified pixel size of 1.06 Å. Each video comprised 50 frames with a total dose of 50 e^-/Å^2 and exposure time was 8 s with a dose rate of 7 e^- per pixel per s on the detector. Data acquisition was done using SerialEM software at −500 nm defocus[33].

Data processing. We collected 2,793 movies and subjected them to motion correction using motioncor2[34]. Contrast transfer function (CTF) estimation was done using Gctf software[35] on the non-dose-weighted micrographs. The particles were picked using gautomatch (developed by K. Zhang, MRC Laboratory of Molecular Biology, Cambridge, UK; http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/). An initial model was made using EMAN2[36] based on a few automatically picked micrographs and using the common-line approach. The particles were extracted in RELION 2.03[37] using a box size of 200 pixels. Picked particles (614,883) were subjected to 3D classification with 5 classes. Particles (190,135) from the best-looking class were subjected to 3D auto-refinement in RELION 2.03. The refined particles were subjected to another run of 3D classification with 5 classes and without alignments, after which 184,672 particles were chosen for a final run of 3D auto-refinement in RELION 2.03. The final map was sharpened with a *B*-factor of −50 Å. Local resolution was determined using RELION[37] with half-reconstructions as input maps. The cryo-EM data collection, refinement and validation statistics are reported in Supplementary Table 1.

Modelling. The initial template for GLP-1R transmembrane regions, G protein and Nb35 was derived from rabbit GLP-1R in complex with $G\alpha_s$ (PDB-5VAI)[2] followed by extensive remodelling using COOT[38]. The ECL3 loop residues 372–376 were stubbed owing to insufficient density for unambiguous modelling, and no high-resolution density was present for ICL3 residues N338–T343, which were omitted from the deposited structure. Owing to discontinuous and/or variable density in the GLP-1R ECD region, we used the high-resolution X-ray crystal structure of the GLP-1R ECD–exendin(9–39) (PDB-3C5T)[16] for a rigid body fit with limited manual adjustments. The ExP5 peptide was modelled manually. The final model was subjected to global refinement and minimization in real space using the module 'phenix.real_space_refine' in PHENIX[39]. Validation was performed in MolProbity[40].

Insect cell membrane preparations. Crude membrane preparations were prepared from insect cells produced using the same expression conditions as used for cryo-EM samples. Cells were resuspended in buffer (20 mM HEPES 7.4, 50 mM NaCl, 2 mM $MgCl_2$, with protease inhibitors and benzonase) and dounced 20 times with the tight pestle, followed by centrifugation (10 min, 350*g*, 4 °C). The pellet was resuspended in buffer, dounced and clarified by centrifugation at a low *g*. Membranes were pelleted by centrifugation (1 h, 40,000*g*, 4 °C), resuspended in buffer and sonicated. Protein concentration was determined using Bradford reagent (Bio-Rad).

[^{35}S]GTPγS binding. Measurement of [^{35}S]GTPγS incorporation was performed in 20 mM HEPES pH 7.4; 100 mM NaCl; 10 mM $MgCl_2$; 1 mM EDTA; 0.1% (w/v) BSA; 30 μg/ml saponin. Membranes (50 μg per sample) were pre-incubated with 1 μM GDP and increasing concentrations of ligand for 30 min at 22 °C. Reactions were started by the addition of [^{35}S]GTPγS and ATP (final concentrations: 300 pM and 50 μM, respectively). After 1 h incubation at 30 °C, the reaction was terminated by collecting the membranes on Whatman UniFilter GF/C plates using Filtermate 196 harvester (Packard). Membranes were extensively washed with ice-cold 50 mM Tris pH 7.6, 10 mM $MgCl_2$, 100 mM NaCl, dried, dissolved in 40 μl MicroScint-O scintillation cocktail (Packard) and counted using a MicroBeta LumiJET counter (PerkinElmer). Data from each experiment were normalized to the response of GLP-1R–WT$G\alpha_s$–$G\beta_1\gamma_2$ membranes at 1 μM GLP-1 (100%).

Radioligand competition binding experiments on insect cell membranes. Radioligand binding was performed in 20 mM HEPES, pH 7.4, 100 mM NaCl, 10 mM $MgCl_2$ and 0.1% (w/v) BSA. Competition binding assays with GLP-1 and ExP5 were performed in the presence of 50 pM [^{125}I]-exendin(9–39). Binding reactions were initiated with the addition of 4 μg of GLP-1R-expressing membranes (with or without G protein) followed by 1 h incubation at 30 °C. Membranes were collected on UniFilter GF/C (Whatman) plates using a Filtermate 196 harvester (Packard), extensively washed with ice-cold NaCl, dried, dissolved in 40 μl of MicroScint-O scintillation cocktail (Packard), and counted using a MicroBeta LumiJET counter (PerkinElmer). Data from each experiment were normalized to vehicle control and non-specific binding (1 μM exendin(9–39)). Curves were fit to a one- or two-site competition binding equation in Prism 6.0 (GraphPad).

Generation of mutant receptor constructs in mammalian cell lines. Mutant receptors were generated in a 2× c-Myc epitope-tagged receptor in the pEF5/FRT/V5-DEST vector using QuikChange site-directed mutagenesis (Invitrogen) and sequences confirmed. Mutant receptors were stably expressed in CHOFlpIn cells using the FlpIn Gateway technology system (Invitrogen) and selected using 600 μg/ml hygromycin B. All cells were tested and found to be free from mycoplasma contamination.

Mammalian whole-cell radioligand binding assays. Cells were seeded at a density of 30,000 cells per well into 96-well culture plates and incubated overnight in DMEM containing 5% FBS at 37 °C in 5% CO_2. Growth medium was replaced with binding buffer (DMEM containing 25 mM HEPES and 0.1% (w/v) BSA) containing 0.1 nM [^{125}I]-exendin(9–39) and increasing concentrations of unlabelled peptide agonists. Cells were incubated overnight at 4 °C, followed by three washes in ice cold 1× PBS to remove unbound radioligand. Cells were then solubilized in 0.1 M NaOH, and radioactivity determined by gamma counting. For all experiments, nonspecific binding was defined by 1 μM exendin(9–39).

Mammalian cAMP assays. Cells were seeded at a density of 30,000 cells per well into 96-well culture plates and incubated overnight in DMEM containing 5% FBS at 37 °C in 5% CO_2. cAMP detection was performed as previously described[3]. All values were converted to cAMP concentration using a cAMP standard curve performed in parallel and data were subsequently normalized to the response of 100 μM forskolin in each cell line.

β-Arrestin recruitment assay. Cells stably expressing GLP-1R–Rluc8 and β-arrestin1–venus were seeded at a density of 30,000 cells per well into 96-well culture plates and incubated overnight in DMEM containing 5% FBS at 37 °C in 5% CO_2. β-Arrestin recruitment was performed as previously described[41].

Mammalian cell membrane preparations for G protein BRET assays. HEK293AΔS-GLP-1R cells were transfected with $G\alpha_s$–venus (inserted at position 72 of $G\alpha_s$ with a GSSSSG linker) or dominant-negative $G\alpha_s$–nanoluc (inserted at position 72 of $G\alpha_s$ with a GSSSSG linker), $G\beta_1$ and $G\gamma_2$–nanoluc or $G\gamma_2$–venus (inserted at the N terminus of $G\gamma_2$ with a GSAGT linker) at a 1:1:1 ratio using PEI. Cell membranes were prepared as described previously[26] and stored at −80 °C. Twenty-four hours after transfection, cells were collected with membrane preparation buffer (20 mM BisTris, pH 7.4, 50 mM NaCl, 1 mM $MgCl_2$, 1× P8340 protease inhibitor cocktail (Sigma-Aldrich), 1 mM DTT and 0.1 mM PMSF). Cells were then homogenized, applied to a stepped sucrose gradient (60%, 40%, homogenate) and centrifuged at 22,500 r.p.m. for 2.5 h at 4 °C. The layers between 40% and homogenate were collected, diluted in membrane preparation buffer and centrifuged at 30,000 r.p.m. for 30 min at 4 °C. The final pellet was resuspended in membrane preparation buffer, and stored at −80 °C. Total protein concentration was determined using a NanoDrop.

G protein conformational determination using BRET. HEK293AΔS cells stably expressing the GLP-1R (tested and confirmed to be free from mycoplasma) were transfected with a 1:1:1 ratio of $G\gamma_2$:nanoluc–$G\alpha_s{}^{72}$:venus–$G\beta_1$ or $G\gamma_2$:venus–dominant-negative $G\alpha_s{}^{72}$:nanoluc–$G\beta_1$ 24 h before collection and preparation of cell plasma membranes (above). Five micrograms per well of cell membrane was incubated with furimazine (1:1,000 dilution from stock) in assay buffer (1× HBSS, 10 mM HEPES, 0.1% (w/v) BSA, 1× P8340 protease inhibitor cocktail, 1 mM DTT and 0.1 mM PMSF, pH 7.4). The GLP-1R-induced BRET signal between $G\alpha_s$ and $G\gamma$ was measured at 30 °C using a PHERAstar (BMG LabTech). Baseline BRET measurements were taken for 2 min before addition of vehicle or ligand. BRET was measured at 15 s intervals for a further 7 min. All assays were performed in a final volume of 100 μl.

Data analysis. Pharmacological data were analysed using Prism 7 (GraphPad). Concentration-dependent response signalling data were analysed as previously described[20] using a three-parameter logistic equation. Signalling bias was quantified by analysis of cAMP accumulation and β-arrestin1 recruitment concentration–response curves using an operational model of agonism modified to directly estimate the ratio of τ/K_A as described previously[5,20,42].

$$Y = \frac{E_{\max} \times (\tau_c/K_A)^n \times [A]^n}{[A]^n \times (\tau_c/K_A)^n + (1 + [A]/K_A)^n}$$

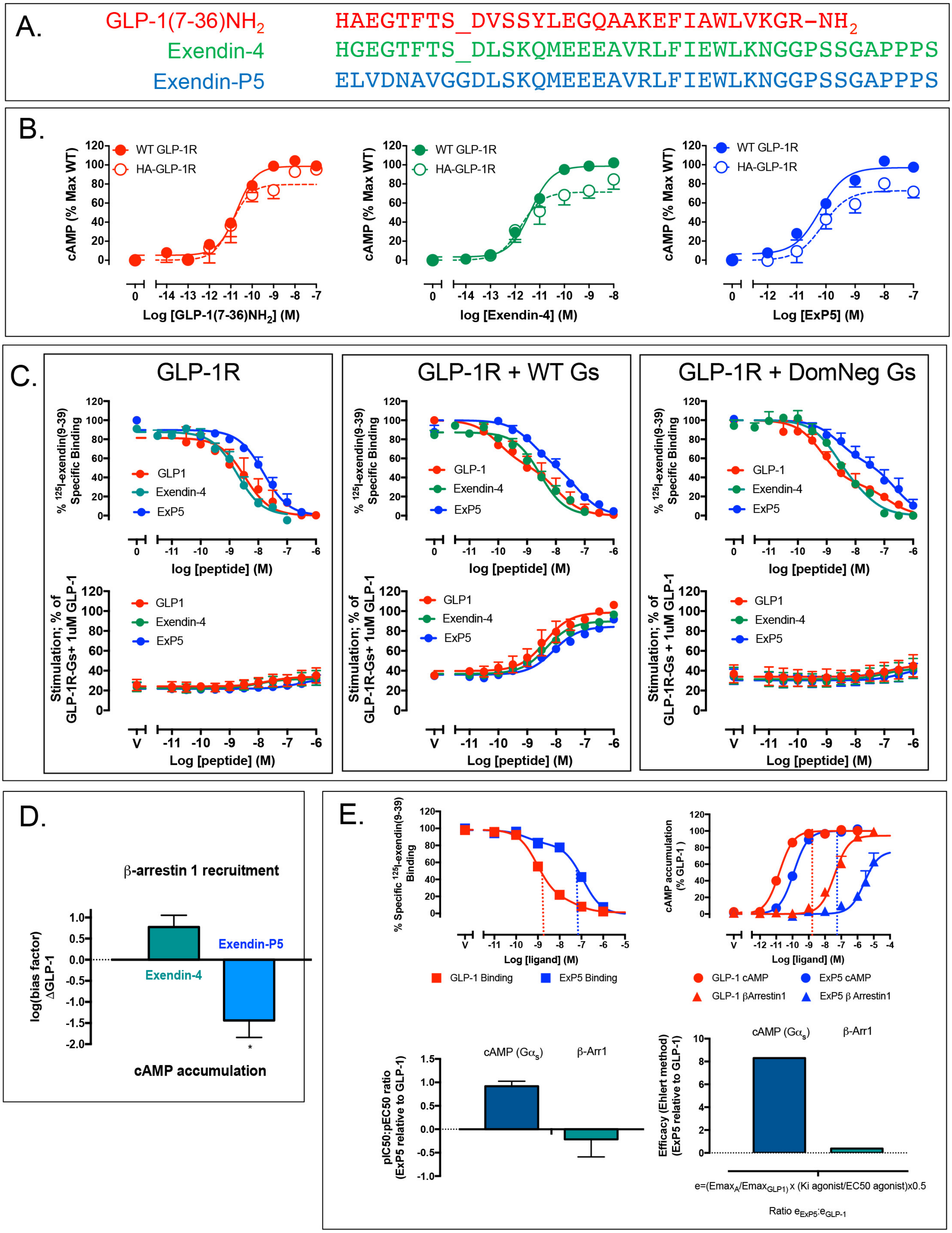
in which $E_{\max}$ represents the maximal stimulation of the system, K_A is the agonist–receptor dissociation constant in molar concentration, $[A]$ is the molar concentration of ligand and τ is the operational measure of efficacy in the system, which incorporates signalling efficacy and receptor density. All estimated τ/K_A ratios included propagation of error for both τ and K_A. Changes in τ/K_A ratios with respect to GLP-1 for each novel peptide were used to quantify bias between signalling pathways. Accordingly, bias factors included propagation of error from the τ/K_A ratios of each pathway.

Changes in the rate of change in BRET signal were fitted to a one-phase association curve. Normalized AUC for the indicated ligand concentrations was plotted as a concentration–response curve and fitted with a three-parameter logistic curve. Statistical analysis was performed with either one-way analysis of variance and a Tukey's post-test or a paired *t*-test, and significance accepted at $P < 0.05$.

Graphics. Molecular graphics images were produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH P41 RR-01081)[43]. Superposition of maps was performed in COOT using "transformation by LSQ model fit"[38]. Measurements of GαRas α5 movements between different structures was performed in Pymol using the psico python module.

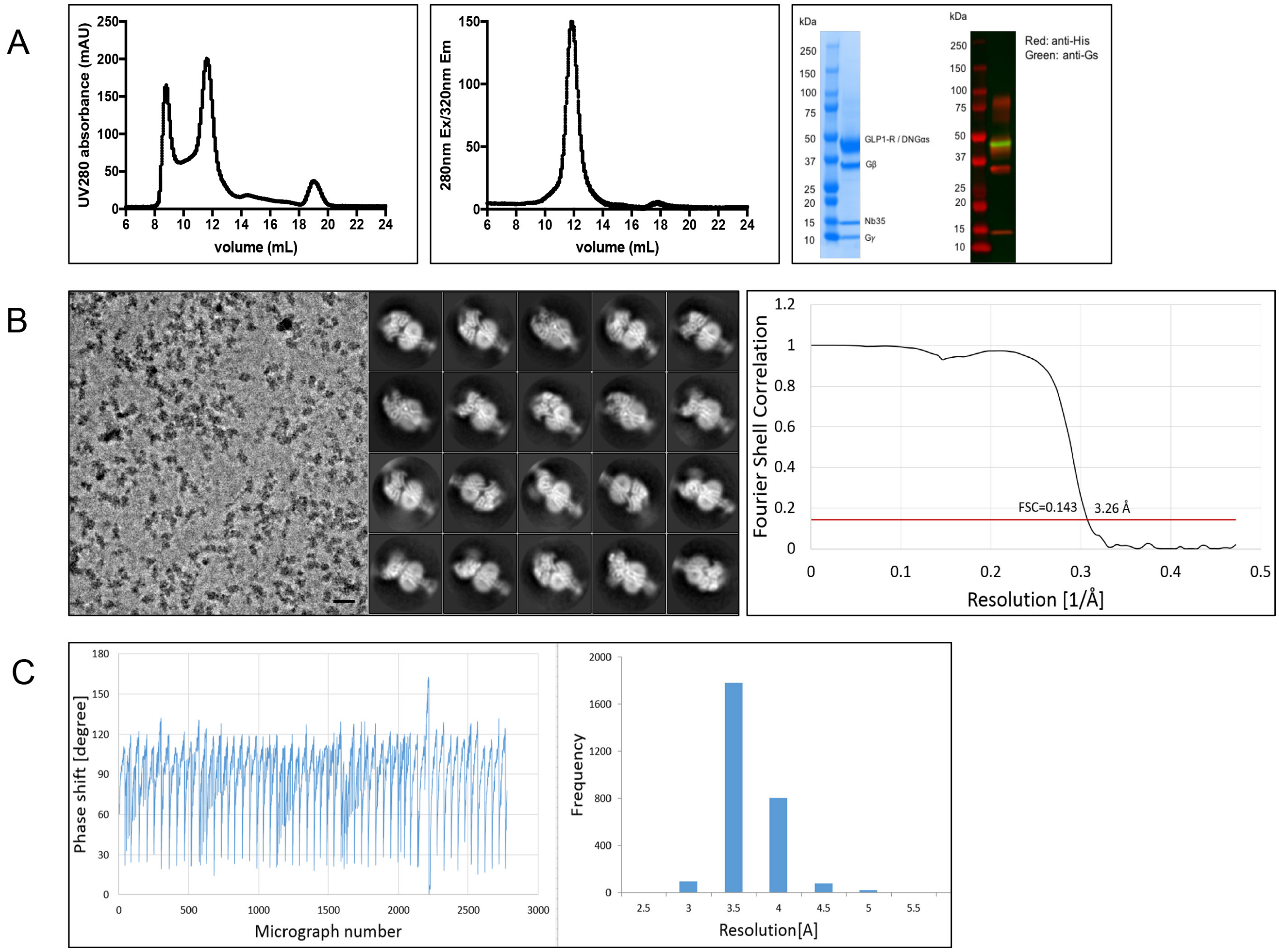
Data availability. All relevant data are available from the authors and/or included in the manuscript or Supplementary Information. Atomic coordinates and the cryo-EM density map have been deposited in the Protein Data Bank (PDB) under accession number 6B3J and EMDB entry ID EMD-7039.

28. Cleator, J. H., Mehta, N. D., Kurtz, D. T. & Hildebrandt, J. D. The N54 mutant of Gαs has a conditional dominant negative phenotype which suppresses hormone-stimulated but not basal cAMP levels. *FEBS Lett.* **443,** 205–208 (1999).
29. Lee, E., Taussig, R. & Gilman, A. G. The G226A mutant of Gsα highlights the requirement for dissociation of G protein subunits. *J. Biol. Chem.* **267,** 1212–1218 (1992).
30. Iiri, T., Bell, S. M., Baranski, T. J., Fujita, T. & Bourne, H. R. A Gsα mutant designed to inhibit receptor signaling through Gs. *Proc. Natl Acad. Sci. USA* **96,** 499–504 (1999).
31. Berlot, C. H. A highly effective dominant negative αs construct containing mutations that affect distinct functions inhibits multiple Gs-coupled receptor signaling pathways. *J. Biol. Chem.* **277,** 21080–21085 (2002).
32. Berlot, C. H. & Bourne, H. R. Identification of effector-activating residues of Gsα. *Cell* **68,** 911–922 (1992).
33. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152,** 36–51 (2005).
34. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14,** 331–332 (2017).
35. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193,** 1–12 (2016).
36. Tang, G.*et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol*. **157,** 38–46 (2007).
37. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5,** e18722 (2016).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
39. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
40. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
41. Savage, E. E., Wootten, D., Christopoulos, A., Sexton, P. M. & Furness, S. G. A simple method to generate stable cell lines for the analysis of transient protein–protein interactions. *Biotechniques* **54,** 217–221 (2013).
42. Hager, M. V. J., Johnson, L. M., Wootten, D., Sexton, P. M. & Gellman, S. H. β-Arrestin-biased agonists of the GLP-1 receptor from β-amino acid residue incorporation into GLP-1 analogues. *J. Am. Chem. Soc.* **138,** 14970–14979 (2016).
43. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
44. Ehlert, F. J. The relationship between muscarinic receptor occupancy and adenylate cyclase inhibition in the rabbit myocardium. *Mol. Pharmacol.* **28,** 410–421 (1985).

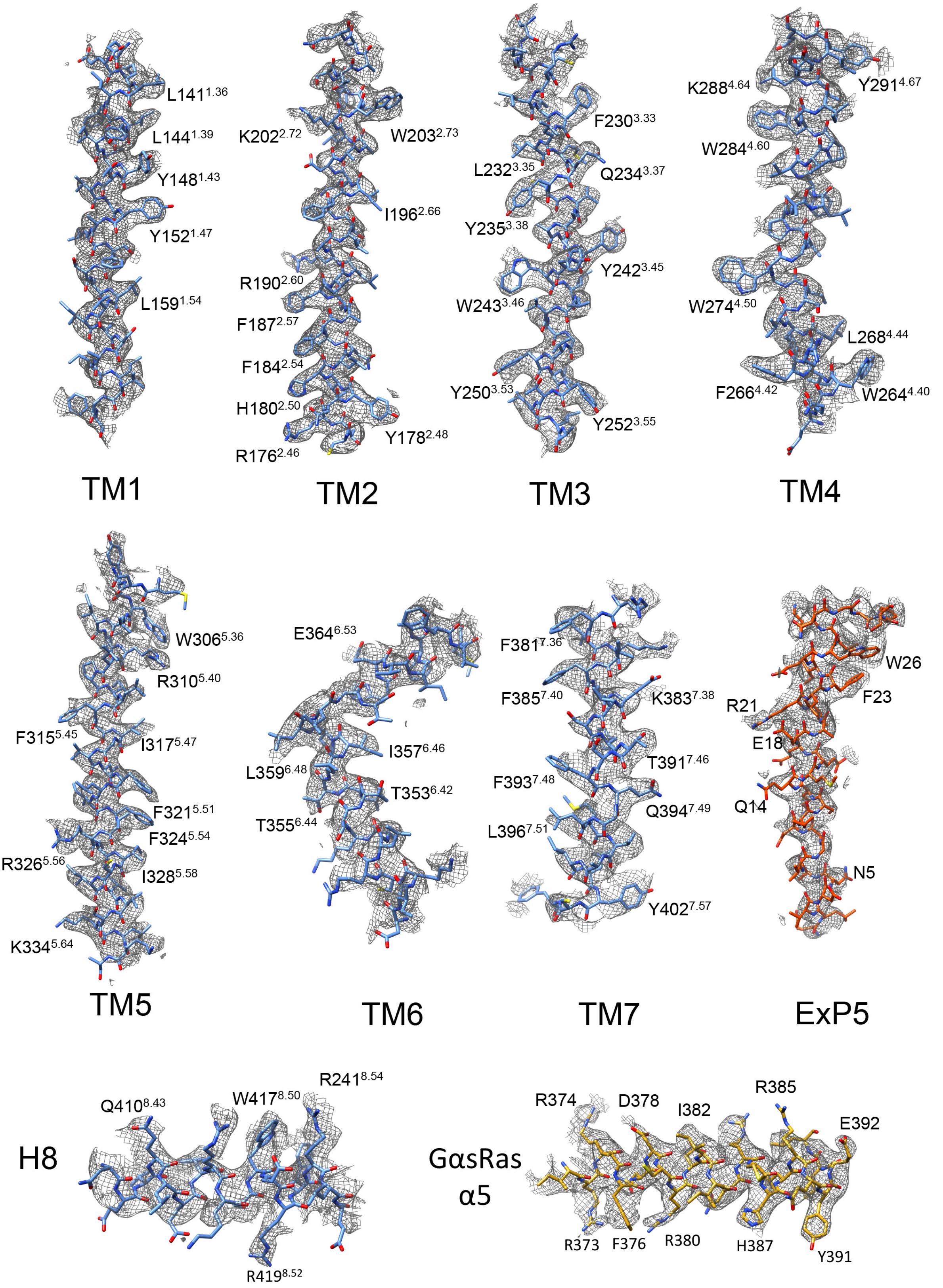


Extended Data Figure 1 | See next page for caption.

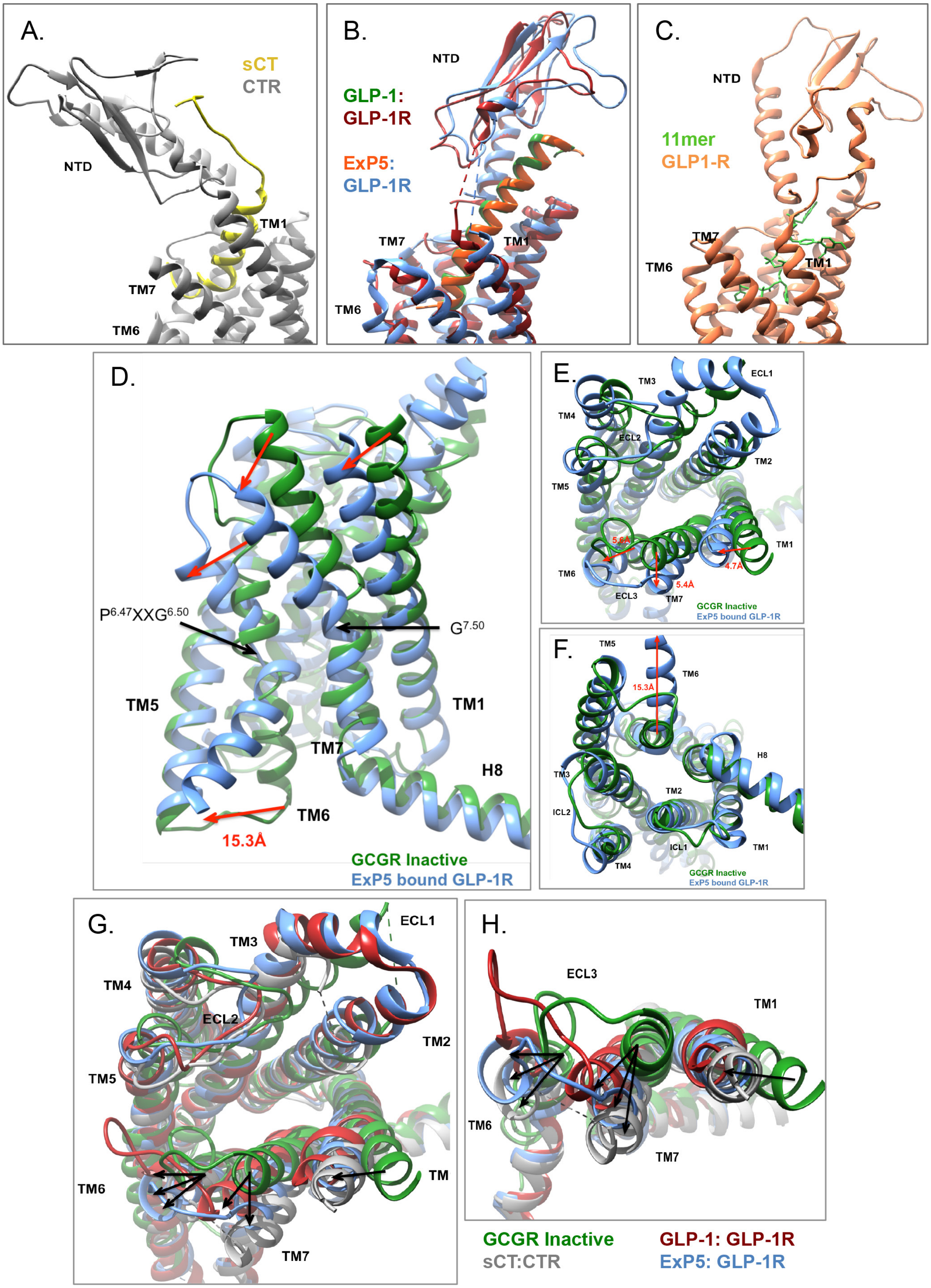
Extended Data Figure 1 | GLP-1R pharmacology. a, Peptide sequences. **b**, Pharmacology of untagged GLP-1R (WT GLP-1R) and the purification construct (HA–GLP-1R). **c**, Insect cell pharmacology of HA–GLP-1R. Top, radioligand competition binding. Bottom, GTPγS binding. Left, no Gs protein. ExP5 has lower affinity than GLP-1 and exendin-4 and does not bind GTPγS. Middle, wild-type Gs enhances peptide affinity and promotes GTPγS binding. Right, dominant-negative Gs is similar to wild-type Gs in binding, but does not bind GTPγS. **d**, Bias factors calculated from concentration–response curves using the Black and Leff operational model[5,20,41] (see Methods) confirm that ExP5 is a biased agonist relative to GLP-1. **e**, Top left, pIC_{50} of ExP5 is ~100-fold lower than of GLP-1 (CHOFlpIn whole cell). Top right, GLP-1 and ExP5 have β-arrestin1 coupling with pEC_{50} ~30-fold to the right of their pIC_{50} (dotted lines). ExP5 is more potent than GLP-1 in cAMP signalling (pEC_{50} relative to pIC_{50}). Bottom left, pIC_{50}:pEC_{50} ratios for G protein (cAMP) and β-arrestin1 of ExP5 relative to GLP-1 highlights ExP5 bias arises from enhanced Gs coupling, not reduced β-arrestin1 recruitment. Bottom right, ratio of ExP5 efficacy (calculated using the Ehlert method[44]) relative to GLP-1 in cAMP and β-arrestin1 recruitment confirms that ExP5 bias arises from enhanced $\mathrm{G}\alpha_s$ efficacy. Data in **b**, **c** are mean ± s.e.m. of three (insect cells) or four (CHOFlpIn cells) independent experiments, conducted in duplicate or triplicate, respectively. Data in **d**, **e** are from 11 independent experiments performed in duplicate. $*P < 0.05$ by one-way analysis of variance and Dunnett's post-test.



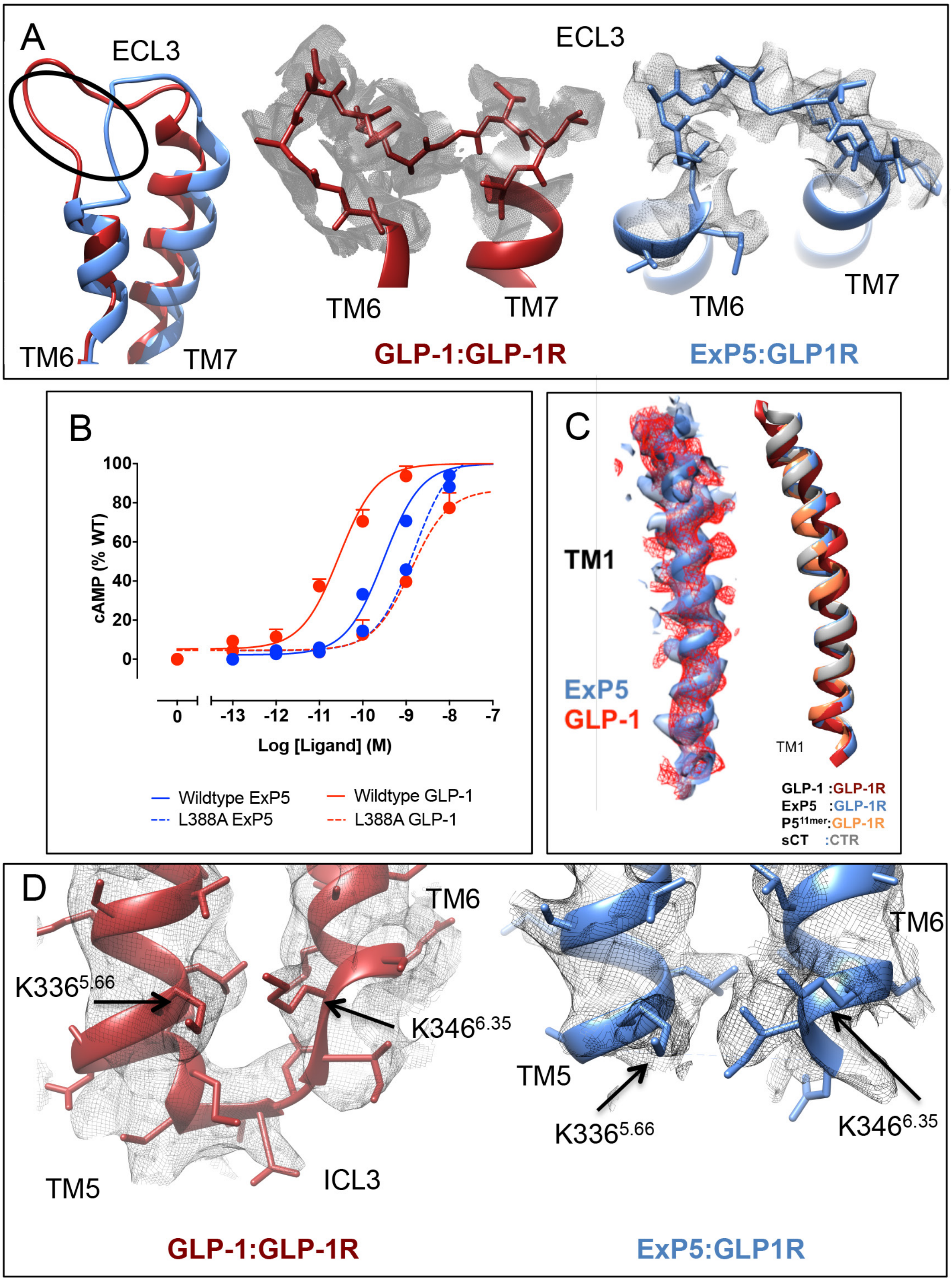
Extended Data Figure 2 | Purification and Volta phase plate imaging of the ExP5–GLP-1R–Gs complex. a, Left, elution profile of the purified complex. Middle, pooled complex fractions, concentrated and analysed by size exclusion chromatography (SEC). Right, SDS–PAGE/Coomassie blue stain and western blot of the complex showing all components. Anti-His antibody detects Flag–GLP-1R–His, Gβ–His and Nb35–His (red) and anti-Gs antibody detects Gα_s (green). **b**, Left, Volta phase plate micrograph of the complex (representative of 2,793). Middle, 2D class averages. Right; 'gold standard' Fourier shell correlation (FSC) curves; the overall nominal resolution is 3.26 Å. **c**, Left, Volta phase plate phase shift history throughout the dataset. Right, histogram of the estimated micrograph resolutions from the CTF.



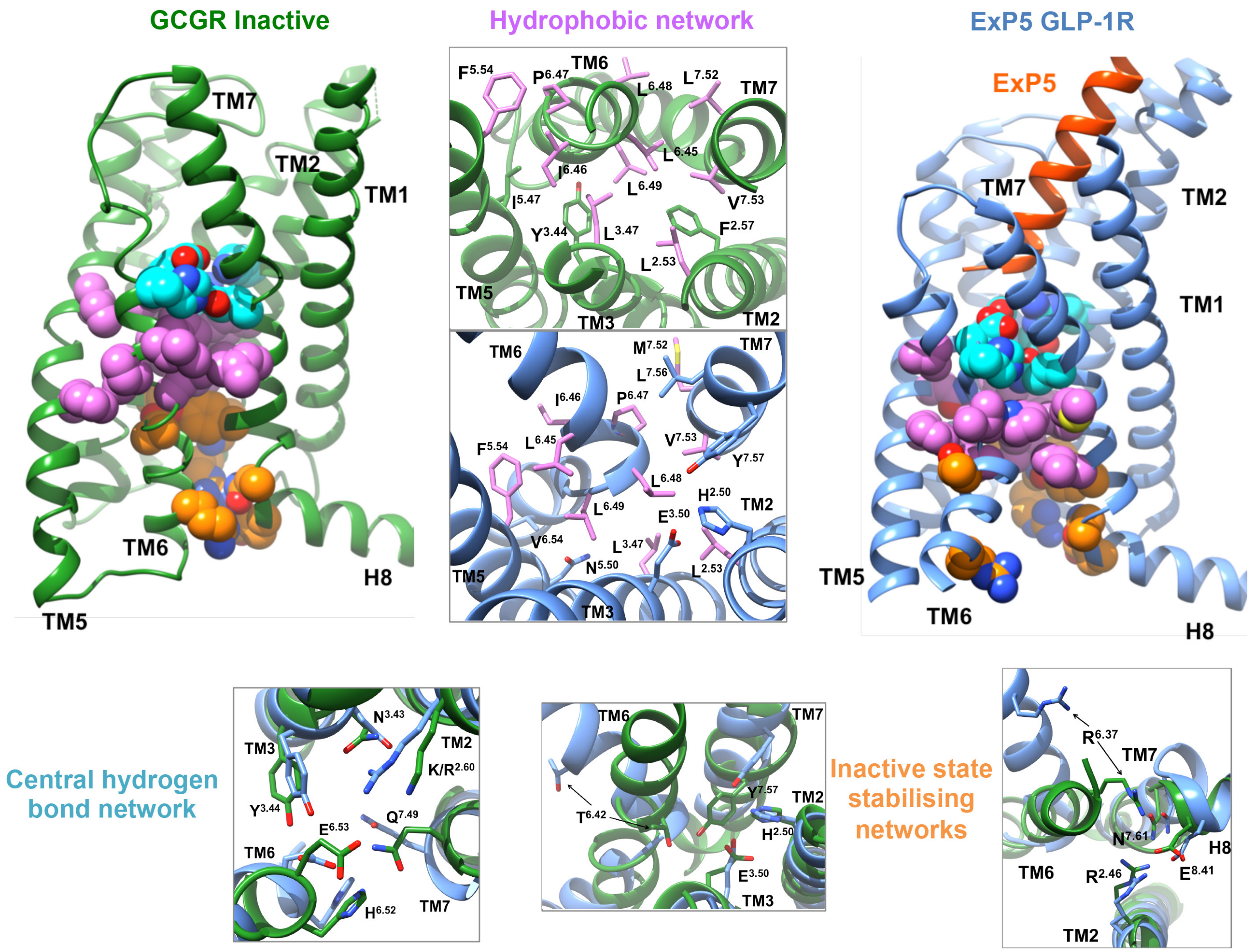
Extended Data Figure 3 | Atomic resolution model of the ExP5–GLP-1R–Gs heterotrimer in the cryo-EM density map. EM density map and model are shown for all seven transmembrane helices and H8 of the receptor, the ExP5 peptide and the α5α helix of the GαSα Ras-like domain. Bulky residues are highlighted. All transmembrane helices exhibit good density, with TM6—which is flexible—being the least well-resolved.



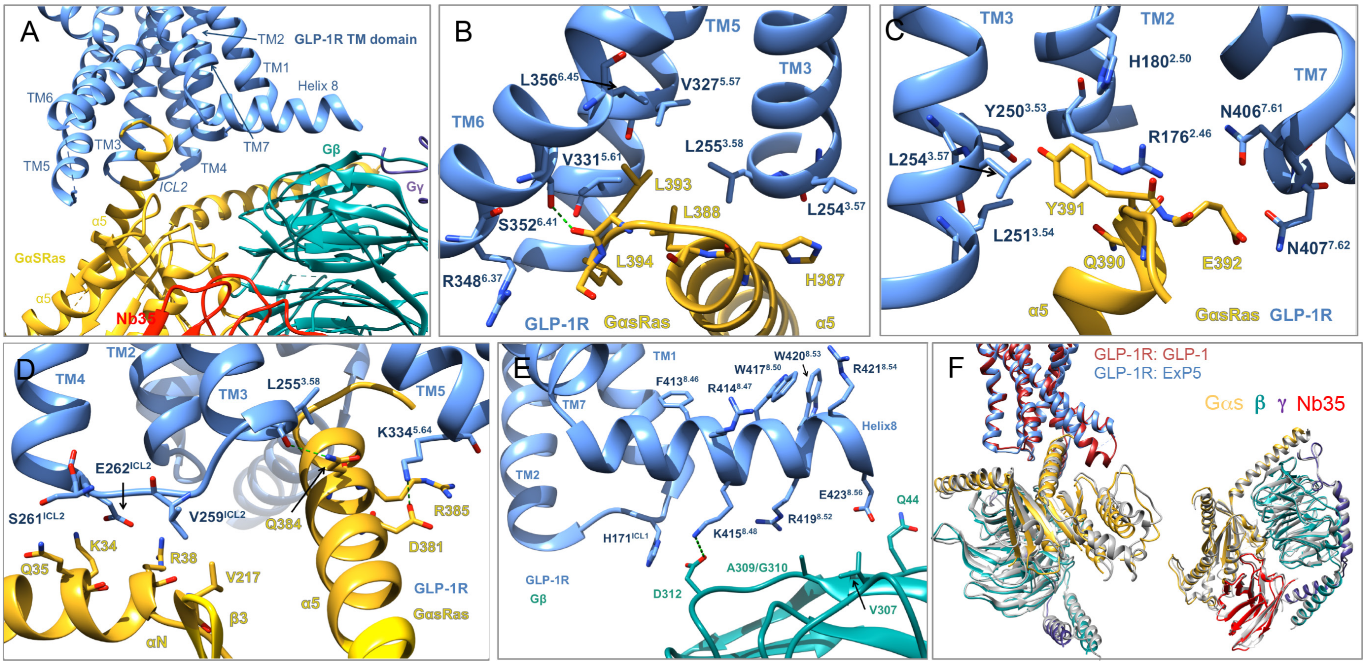
Extended Data Figure 4 | Comparison of class B GPCR structures. **a**–**c**, Agonist-bound full-length structures have distinct NTD orientations. **d**–**f**, Side view (**d**), extracellular view (**e**) and cytoplasmic view (**f**) of the conformational reorganization between inactive (GCGR, PDB 4L6R) and active structures (ExP5-bound GLP-1R). Distances are measured from Cα residues 1.33, 6.58, 7.35 and 6.35. Numbering uses the Wootten class B system. **g**–**h**, Superimposition of transmembrane domains from sCT–CTR–Gs (grey, PDB 5U27), GLP-1–GLP-1R–Gs (red, PDB 5VAI) and ExP5–GLP-1R–Gs with the inactive GCGR (green, PDB 4L6R). The largest differences in active structures relative to the inactive GCGR occur in TM1, TM6, TM7 and ECL3 (**h**), but the nature and extent of conformational change varies.



Extended Data Figure 5 | ECL3, TM7, TM1 and ICL3 may be associated with GLP-1R biased agonism. a, Conformational differences in GLP-1R ECL3 between ExP5-bound (blue) and GLP-1-bound (red) GLP-1R structures are supported by density in their respective cryo-EM maps. **b**, L388$^{7.43}$A affected the potency of GLP-1 mediated cAMP more than ExP5 (mean + s.e.m. of four independent experiments). **c**, Right, TM1 overlays from agonist-bound class B GPCR structures reveals a different conformation for GLP-1–GLP-1R. Left, TM1 model overlays of ExP5–GLP-1R and GLP-1–GLP-1R with their associated cryo-EM maps (GLP-1, red ribbon/mesh; ExP5, blue ribbon/surface) reveals limited differences in the TM1 backbone, but potentially distinct side-chain orientations. **d**, Left, ICL3 backbone conformation in GLP-1–GLP-1R (PDB 5VAI) is supported by density (EMD-3653). Limited density is observed for ICL3 (337–343) in ExP5–GLP-1R.



Extended Data Figure 6 | Rearrangement of conserved networks upon GLP-1R binding to ExP5. Comparison of conserved networks in the inactive (green, GCGR) and activated (blue, ExP5–GLP-1R–Gs) states; central polar network (cyan), cytoplasmic polar networks (orange) and hydrophobic residues (pink). Inactive state interactions are incompatible with peptide binding and reorganize on activation. Upper middle, major rearrangements within the hydrophobic network (top, inactive; bottom, activated); side chains involved in ground state stabilization in green, inactive and active state in pink and active state in blue. Lower left and lower right, reorganization of the central hydrogen bond network and cytoplasmic networks, respectively, where green is inactive and blue is active. Subscript, Wootten numbering. These conformational changes are detailed in Supplementary Video 1.



Extended Data Figure 7 | GLP-1R–G protein interactions. a, GLP-1R forms interactions with $G\alpha_S$Ras and Gβ. **b**–**e**, Receptor side chains (blue) within 4.5 Å of $G\alpha_S$ side chains (gold) or Gβ side chains (cyan). **b**–**d**, $G\alpha_S\alpha5$ forms polar and non-polar interactions with the cytoplasmic cavity formed by TM6 opening. Potential interactions also occur between $G\alpha_S\alpha N$ and ICL2 of GLP-1R. **e**, GLP-1R H8 aromatic residues embed within the detergent micelle and polar residues form direct interactions with Gβ. **f**, Left, the distinct engagement angle of $G\alpha_S\alpha5$ with the receptor (Fig. 3) results in an overall rotation of the $G\alpha_s$Ras,β,γ in ExP5–GLP-1R relative to GLP-1–GLP-1R. Right, overlaying $G\alpha_s$ from both structures reveals only minor differences in the G protein upon receptor engagement.

Extended Data Table 1 | Effects of extracellular loop 3 alanine mutants of human GLP-1R on agonist binding and cell surface expression

GLP-1R	Cell surface expression	Whole cell competition radioligand binding pK_i			cAMP Accumulation					
		GLP-1	Exendin-4	Exendin-P5	GLP-1		Exendin-4		Exendin-P5	
					pEC_{50}	E_{max}	pEC_{50}	E_{max}	pEC_{50}	E_{max}
Wild type	100 ± 3	8.81±0.04 -	8.97±0.07 -	9.61±0.45(0.18) 7.05±0.09	10.8±0.04	100±2	11.4±0.04	100±1	9.93±0.04	100±1
D372A	86 ± 4	9.62±0.47 (0.15)	9.29±0.56 (0.14) **7.40±0.21***	9.34±0.41 (0.24) 6.85±0.09	**8.95±0.10***	103±4	**10.1±0.09***	105±3	10.0±0.08	99±2
E373A	94 ± 7	9.80±0.39 (0.26)	9.67±0.42 (0.10) **7.41±0.15***	9.55±0.29 (0.22) 6.71±0.08	**8.45±0.08***	104±4	**9.3±0.09***	99±3	**9.25±0.08***	106±3
H374A	82 ± 8	8.87±0.14 -	8.87±0.09 -	9.94±0.52 (0.10) 7.24±0.13	10.9±0.07	102±2	11.6±0.08	99±2	10.1±0.07	93±2
R376A	92 ± 7	8.53±0.10 -	8.92±0.10 -	9.61±0.48 (0.15) 7.05±0.09	**10.3±0.09***	99±3	11.4±0.12	95±2	9.61±0.14	105±5
G377A	96 ± 3	**8.27±0.06***	9.12±0.25 (0.68) **7.29±0.21***	- 7.03±0.10	**9.9±0.08***	105±3	11.1±0.15	102±4	**9.14±0.15***	106±5
T378A	173 ± 11*	8.87±0.06	8.71±0.11	- 6.75±0.08	10.7±0.10	93±2	11.5±0.17	85±7	10.0±0.10	95±3
L379A	90 ± 6	**8.33±0.07***	**8.15±0.15***	9.61±0.49 (0.15) **6.63±0.11***	**9.74±0.13***	89±3	**10.5±0.20***	92±5	10.2±0.10	96±3
R380A	73 ± 5	**7.35±0.09***	**7.65±0.05***	**-** **6.65±0.04***	**7.74±0.07***	103±4	**8.81±0.08***	103±3	**8.21±0.11***	97±4
F381A	92 ± 5	8.97±0.06	8.94±0.05	9.22±0.51 (0.19) **6.32±0.15***	10.5±0.06	88±2	11.5±0.08	86±6	10.3±0.06	93±2
I382A	115 ± 6	8.92±0.06	8.91±0.06	9.71±0.51 (0.14) 6.79±0.12	10.9±0.12	99±3	11.3±0.09	95±2	10.2±0.12	102±4

Cell surface expression was determined through antibody detection of the N-terminal c-Myc epitope label and expressed as percentage of wild-type (WT) GLP-1R expression. Whole-cell competition radioligand binding data were analysed using either a one-site (a single pK_i) or a two-site binding curve (two pKi values are reported with the fraction of receptors in the high affinity site reported in brackets) as determined by an *F*-test in Graphpad Prism. pK_i values represent the negative logarithm of the equilibrium dissociation constant (in molar) of agonist. Data were normalized to specific [^{125}I]-exendin(-9-39) binding. cAMP concentration response data were analysed using a three-parameter logistic curve to determine pEC_{50} and E_{max} values. pEC_{50} values represent the negative logarithm of agonist concentration that produces half maximal response. E_{max} values are maximal response as percentage of WT response. All values are expressed as mean ± s.e.m. of five independent experiments conducted in duplicate. Data were analysed using one-way analysis of variance and Dunnett's post-test. $*P < 0.05$ (in comparison with WT response).

Extended Data Table 2 | Interactions between the GLP-1R and ExP5

ExP5	Peptide side chain density at Cβ	Peptide side chain density at Cγ	GLP-1R	Interaction
E1	yes	no	R310$^{5.40}$	Hydrogen bond
			A368$^{6.57}$	
L2	yes	yes	V237$^{3.40}$	
			I313$^{5.43}$	
V3	yes	N/A	L384$^{7.39}$	
			E387$^{7.42}$	
			L388$^{7.43}$	
			T391$^{7.46}$	
D4	no	no	Y152$^{1.47}$	Potential H-bond
			V194$^{2.64}$	
			M233$^{3.36}$	
			K197$^{2.67}$	Salt bridge
N5	yes	yes	Q234$^{3.37}$	Hydrogen bond
			W306$^{5.36}$	Hydrogen bond
A6	yes	N/A		
V7	yes	N/A	L384$^{7.39}$	
			L388$^{7.43}$	
G8	N/A	N/A		
G9	N/A	N/A		
D10	yes	no	R380$^{7.35}$	Salt bridge
L11	yes	no	L141$^{1.36}$	
			Y145$^{1.40}$	
			L201$^{2.71}$	
S12	yes	N/A	T298^{ECL2}	
			L201$^{2.71}$	
K13	yes	yes	R299^{ECL2} backbone	potential H-bond to the backbone
Q14	yes	yes	E138$^{1.33}$	no side chain density for E138
			L141$^{1.36}$	
M15	yes	yes	L201$^{2.71}$	
			K202$^{2.71}$	
			Y205$^{2.75}$	
			S206$^{2.76}$	
E16	yes	yes	Y205$^{2.75}$	Potential H-bond
			R299^{ECL2}	Salt bridge, Potential H-bond to the backbone
			N32	
E17	yes	no	*Potentially TM1 stalk*	
E18	yes	yes	*Potentially TM1 stalk*	
A19	yes	N/A	Y205$^{2.75}$	
V20	yes	N/A	*V30*^{NTD}	
			L32^{NTD}	
			P90^{NTD}	
R21	yes	yes	*Potentially TM1 stalk*	
L22	yes	yes	*Q210*^{ECL1}	
F23	yes	yes	W214^{ECL1}	
			L32^{NTD}	
			L35^{NTD}	
			V36^{NTD}	
			W39^{NTD}	
I24	yes	yes	V20 ^{NTD}	
			Y69^{NTD}	
			L89^{NTD}	
			P90^{NTD}	
			W91^{NTD}	
E25	yes	yes		
W26	yes	yes	H212^{ECL1}	
			W214^{ECL1}	π-stack
L27-S33			N-terminal interactions	

Residues in black are within 4 Å of the bound peptide. Residues in grey italics are within 4.5 Å of the bound peptide, but out of bonding distance and may form transient interactions. Residues in blue italics are within 4 Å in our model but there is no side-chain density in the cryo-EM map.

Extended Data Table 3 | Interactions formed between class B receptor and Gs heterotrimeric Gs proteins

G protein subunit	G protein Residue no	GLP-1R (ExP5 bound)	GLP-1R (GLP-1 bound) PDB: 5VAI	CTR (sCT bound) PDB:5UZ7
GαRas α5	R380	*F257*^{ICL2}	L256$^{2.59}$bb	K326$^{6.54}$
	D381	K334$^{5.64}$	K334$^{5.64}$	
	I383	*S258*^{ICL2}		V252^{ICL2}
	Q384	L255$^{2.58}$bb K334$^{5.64}$	L255$^{2.58}$bb K334$^{5.64}$	L248$^{2.58}$bb K326$^{5.64}$
	R385	K334$^{5.64}$ bb	K334$^{5.64}$bb	K326$^{5.64}$
	H387	L254$^{3.57}$ L255$^{2.58}$	L254$^{3.57}$	L247$^{3.57}$
	L388	L255$^{2.58}$ V331$^{5.61}$ *K334*$^{5.64}$	V331$^{5.61}$	L323$^{5.61}$
	Q390	R176$^{2.46}$	R176$^{2.46}$ E408$^{8.41}$	R180$^{2.46}$
	Y391	*H180*$^{2.50}$ Y250$^{3.53}$ L251$^{3.54}$ ***L254***$^{3.57}$	H180$^{2.50}$ L359$^{6.48}$ L356$^{6.45}$ L251$^{3.54}$	Y253^{ICL2} L244$^{3.54}$
	E392	*N406*$^{7.61}$ N407$^{7.62}$	N406$^{7.61}$ V405$^{7.60}$bb L401$^{7.56}$bb	***C394***$^{7.60}$***bb*** ***N396***$^{7.62}$
	L393	S352$^{6.41}$bb L356$^{6.45}$ ***V327***$^{5.57}$ *V331*$^{5.61}$	***S352***$^{6.41}$ L356$^{6.45}$ T353$^{6.42}$	L348$^{6.45}$
	L394	*V331*$^{5.61}$ *R348*$^{6.37}$ *(to L394 backbone)*	*L339*$^{5.59}$	M327$^{5.65}$
GαRas αN	Q35	S261^{ICL2}	S261^{ICL2} E262^{ICL2}	
	R38	E262^{ICL2}		
	Q31	Q263^{ICL2} (not support density)	Q263^{ICL2}	
	K34	*E262*^{ICL2}	Q263^{ICL2}	
GαRas β3	V217	V259^{ICL2}		
GαRas α4	R385	*(N338 not resolved in cryo EM map but likely conserved)*	N338^{ICL3}	
Gβ	D312	H171^{ICL1} K415$^{8.48}$	H171^{ICL1} K415$^{8.48}$	R404$^{8.48}$
	A309/G310 (backbone)	R419$^{8.52}$	R419$^{8.52}$	Q408$^{8.52}$
	Q44	*E423*$^{8.56}$		Q415$^{8.60}$

All receptor residues within 4 Å (4.5 Å in non-bold italics) of G protein that were evident in the cryo-EM maps of the sCT–CTR–Gs, GLP-1–GLP-1R–Gs and ExP5–GLP-1R–Gs complexes are listed. Residues in red are conserved interactions between the three structures, those in blue are conserved between the two GLP-1R structures and those in black are unique in the different structures (bb indicates backbone interactions).

TOOLBOX

THE MAZES WITH MINDS OF THEIR OWN

Automated 'smart mazes' free behavioural researchers from the tedium of monitoring animals. They also boost data quality and reproducibility.

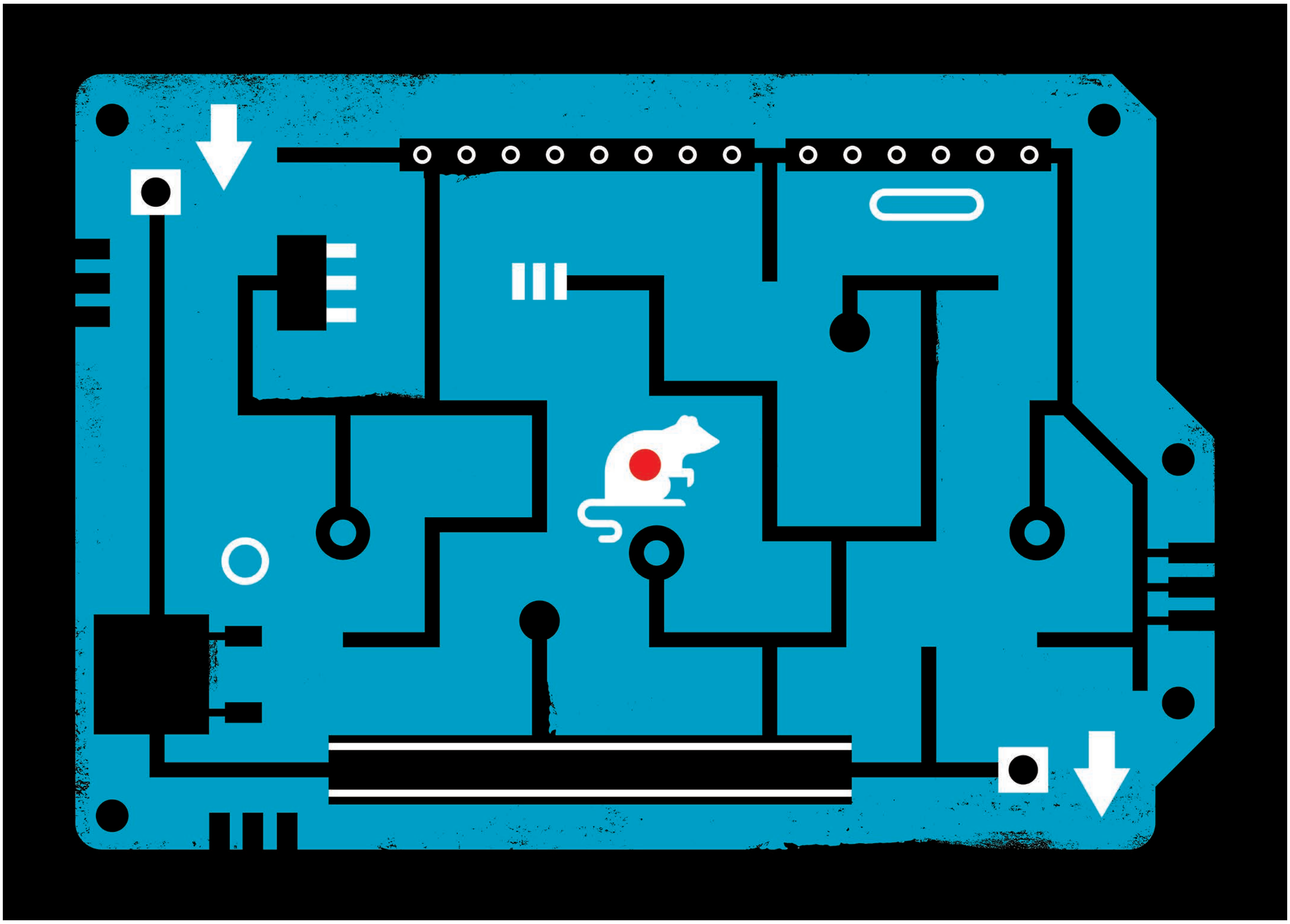


ILLUSTRATION BY THE PROJECT TWINS

BY CHARLES Q. CHOI

As a graduate student, behavioural neuroscientist Mark Brandon spent hours running rodents through a T maze, a test of learning in which animals run down a track and then turn either right or left. The task was useful, but boring. So, when he secured a faculty position at McGill University in Montreal, Canada, Brandon wanted to skip the monotony.

He acquired an automated T maze from MazeEngineers, a start-up firm in Cambridge, Massachusetts. Much more than a simple labyrinth, the maze has doors that rise from the floor after a rodent passes to stop it going backwards, and integrated video monitoring to document the animal's behaviour. Once the rodent has completed a task, the maze directs it back to the beginning. By analysing which neural circuits are active during these tests, Brandon and his colleagues hope to shed light on how the brain links memories with time. "The automated T maze has been incredibly helpful," he says.

Today, such systems are becoming increasingly common and sophisticated. "We're starting to see technologies that have made major advances elsewhere, like touchscreens and microcontrollers, make their way into behavioural-research labs," says Alexxai Kravitz, a neuroscientist at the US National Institute of Diabetes and Digestive and Kidney Diseases ▶

▶ in Bethesda, Maryland. But despite the automation, the human touch is often still needed to ensure these gadgets work correctly.

Researchers use mazes to test an animal's mental traits, such as memory and attention. For much of the twentieth century, mazes forced rodents down elaborate paths, exploring mental limits. Today's set-ups are considerably less convoluted, says MazeEngineers founder Shuhan He, making the results easier to incorporate into statistical analyses. "Mazes to me are about decisions and outcomes — they allow key tests of cognition." But because researchers increasingly want to look at lots of animals in each experiment and to collect as many behavioural data as possible, automation has become a necessity.

Scientists first used video tracking to monitor rodent behaviour automatically in the 1990s, starting with computers that identified an animal's centre of mass to recognize simple behaviours, such as whether the rodent sought out or avoided a particular stimulus. Now, by tracking the head and other body parts, software can recognize subtler behaviours, such as sniffing, grooming and urinating. Tracking systems also exist for creatures such as fruit flies and zebrafish.

"Now you can just start a program and automatically track an animal in a maze, and get the data to analyse instantly after the animal is finished," says Justin Rhodes, a behavioural neuroscientist at the University of Illinois in Urbana–Champaign who studies the effects of exercise on the brain.

Automation also promises increased accuracy and objectivity by reducing the human role in experiments — for instance, when it comes to deciding whether an animal froze or merely paused. And it provides standardized equipment and tasks so that researchers can compare their results directly.

MazeEngineers' automated T mazes for mice start at US$4,900. Video-based mazes and software are also available from such companies as Clever Sys in Reston, Virginia, and Noldus in Wageningen, the Netherlands. Rhodes favours Clever Sys's HomeCageScan system (about $55,000 for four cages) for its ability to track multiple animals. That, he says, is not trivial, "given how they can climb over and under each other". Behavioural neuroscientist Evelin Cotella at the University of Cincinnati in Ohio prefers Noldus's EthoVision software ($5,850 per licence), for its user-friendly interface.

That's not to say that smart-maze systems work perfectly out of the box. Scientists often have to fiddle with settings including lighting and visual contrast to get video-tracking software to work, Rhodes says. And they must measure how accurate their systems are at recognizing behaviours. "It's not going to be perfect, but people aren't, either," Cotella says. "What software will be, compared to people, is consistent."

Cotella's team had to test how well its system categorized behaviour, compared with the researchers. The team also had to optimize parameters such as how many pixels per video frame had to remain unchanged to count as the mouse 'freezing'. It took about six months for the group to feel confident with the system, says Cotella. "We've now had it for a year, and we'll start publishing studies from it in the next month. It's definitely helped us move faster."

Another option, the IntelliCage from TSE Systems in Bad Homburg, Germany, tracks animal movement using subcutaneous radio-frequency identification (RFID) tags. "You don't have to worry about light like you do with video tracking; it runs even in total darkness," explains Daniela Oettler, a scientific associate at TSE.

IntelliCage fits inside any conventional lab cage. Each of its four corners houses devices that can run rodents through some of the tests they might experience in mazes — for example, they might receive an unpleasant puff of air in one place, but not others. Each corner can also be keyed to respond only to specific animals, so that different rodents undergo different tests.

At $60,000, the IntelliCage system holds 16 mice. The ability to keep multiple mice in one cage is a plus, notes behavioural neuroscientist David Wolfer of the University of Zurich in Switzerland, because they are social animals. "Housing them together helps lower stress, which is a source of variability in the data." That said, because it is based on RFID tags, IntelliCage recognizes fewer behaviours than video, Wolfer says. For instance, it cannot detect when an animal rears up on its hind legs. But Oettler counters that video analysis is more open to observer bias in terms of interpreting results.

WATCH WHAT YOU EAT

Behavioural scientists have also automated dietary monitoring. Kravitz, for instance, investigates obesity by tracking mouse food intake and activity levels. But because mice eat very little, even tiny mistakes in measurements can throw results off, he says. That's where automation comes into play.

Automated hardware from firms such as Research Diets in New Brunswick, New Jersey, and Sable Systems International in North Las Vegas, Nevada, uses miniature electronic scales to measure each animal's consumption. But it can have a high price-tag, making studies of many animals at once economically impractical. So Kravitz and his co-workers developed an open-source alternative: the Feeding Experimentation Device (FED).

Designed to fit in a standard cage, FED uses an animal's RFID tag to document its behaviour, logging every time a mouse takes a food pellet. Each device costs roughly $350 to build, less than one-tenth of the price of commercially available systems.

Instructions are available on OpenBehavior, a site co-founded by Kravitz that is dedicated to open-source behavioural-science projects. But, cautions Kravitz, do-it-yourselfers are usually on their own if the system has technical hiccups. "You build them yourself. They're cheap to build, but how cheap depends on how much your time is worth," he says.

RODENT IPADS

At Western University in London, Canada, cognitive neuroscientists Tim Bussey and Lisa Saksida have developed chambers containing touchscreens, which researchers can use to test rodents on 20 or so cognitive tasks, covering memory, learning, attention and even gambling.

"It's like having an iPad for a mouse or rat," Bussey says. "The main difference is that where a human would touch their finger to the screen, a mouse would bring their nose to their screen." Successful completion of a task causes a built-in food-well to dispense drops of strawberry milkshake.

The touchscreens come preprogrammed with tasks from Campden Instruments in Loughborough, UK, and Lafayette Instrument in Indiana. Alternatively, researchers can use software called the Animal Behavior Environment Test System to program their own tasks, says Bussey.

Because the tasks are standardized, data can be compared across labs, or even species, says Brandon, who uses the devices in his research. The aim is to place information from many labs in one database so that the community can tell what neurons are doing in structures across the brain during certain tasks. "The amount of data we can collect blows my mind."

And that volume of data is growing. Brandon is now coupling touchscreen chambers with open-source 'miniscopes' that use fluorescence imaging to record neural activity in freely moving mice for about $500 per microscope. The chambers can even be integrated with optogenetics, experiments that use light to manipulate neural activity, he says.

And some firms are pushing automation even further, with fully automated habitats that allow around-the-clock monitoring, for instance to expand knowledge of the effects of a drug. TSE's PhenoWorld system, for example, houses several RFID-tagged rats or mice in multiple arenas, mazes and floors for experiments. It simultaneously records metabolic and other traits of the animals; a system for a whole colony might cost $800,000, says Oettler.

Similarly, researchers at MazeEngineers are building the 'Labyrinth', a grid of up to 25 modules that can be configured into more than 20 automated mazes. Housing units for rodents could be attached to the Labyrinth's edges, and automatically let animals in and out. "My dream is that the Labyrinth can run by itself," says He.

If that becomes reality, mazes might go from being tools of science to acting like scientists themselves. ■

Charles Q. Choi *is a freelance science writer in New York City.*

SPEAKING OUT Former government scientist raises climate megaphone **p.131**

SCIENCE ON CAMERA Send us your best research pictures for our contest **p.131**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM ARON VELLEKOOP LEON/GETTY

PUBLISHING

The write stuff

How to produce a first-class paper that will get published, stand out from the crowd and pull in plenty of readers.

Manuscripts may have a rigidly defined structure, but there's still room to tell a compelling story — one that clearly communicates the science and is a pleasure to read. Scientist-authors and editors debate the importance and meaning of creativity and offer tips on how to write a top paper.

ANGEL BORJA
Keep your message clear

Marine scientist at AZTI-Tecnalia, a producer of sustainable business services and goods, Pasaia, Spain; journal editor; author of a series on preparing a manuscript (go.nature.com/2gu4hp9).

Think about the message you want to give to readers. If that is not clear, misinterpretations may arise later. And a clear message is even more important when there is a multidisciplinary group of authors, which is increasingly common. I encourage groups to sit together in person and seek consensus — not only in the main message, but also in the selection of data, the visual presentation and the information necessary to transmit a strong message.

The most important information should be in the main text. To avoid distraction, writers should put additional data in the supplementary material.

Countless manuscripts are rejected because the discussion section is so weak that it's obvious the writer does not clearly understand the existing literature. Writers should put their results into a global context to demonstrate what makes those results significant or original.

There is a narrow line between speculation and evidence-based conclusions. A writer can speculate in the discussion — but not too much. When the discussion is all speculation, it's no good because it is not rooted in the author's experience. In the conclusion, include a one- or two-sentence statement on the research you plan to do in the future and on what else needs to be explored.

DALLAS MURPHY
State your case with confidence

Book author, New York City; instructor, writing workshops for scientists in Germany, Norway and the United States.

Clarity is the sole obligation of the science writer, yet I find constantly that the 'What's new' element is buried. Answering one central question — What did you do? — is the key to finding the structure of a piece. Every section of the manuscript needs to support that one fundamental idea.

There is a German concept known as the 'red thread', which is the straight line that the audience follows from the introduction to the conclusion. In science, 'What's new and compelling?' is the red thread. It's the whole reason for writing the paper. Then, once that's established, the paragraphs that follow become the units of logic that comprise the red thread.

Scientific authors are often scared to make confident statements with muscularity. The result is turgid or obfuscatory writing that sounds defensive, with too many caveats and long lists — as if the authors are writing to fend off criticism that hasn't been made yet. When they write for a journal gatekeeper rather than for a human being, the result is muddy prose.

Examples such as this are not uncommon: "Though not inclusive, this paper provides a useful review of the well-known methods of physical oceanography using as examples various research that illustrates the methodological challenges that give rise to successful solutions to the difficulties inherent in ▶

▸ oceanographic research." Why not this instead: "We review methods of oceanographic research with examples that reveal specific challenges and solutions"?

And if the prose muddies the science, the writer has not only failed to convey their idea, but they've also made the reader work so hard that they have alienated him or her. The reader's job is to pay attention and remember what they read. The writer's job is to make those two things easy to do. I encourage scientists to read outside their field to better appreciate the craft and principles of writing.

ZOE DOUBLEDAY

Beware the curse of 'zombie nouns'

Ecologist, University of Adelaide, Australia; co-author of a paper on embracing creativity and writing accessible prose in scientific publications.

Always think of your busy, tired reader when you write your paper — and try to deliver a paper that you would enjoy reading yourself.

Why does scientific writing have to be stodgy, dry and abstract? Humans are story-telling animals. If we don't engage that aspect of ourselves, it's hard to absorb the meaning of what we're reading. Scientific writing should be factual, concise and evidence-based, but that doesn't mean it can't also be creative — told in a voice that is original — and engaging (Z. Doubleday *et al. Trends Ecol. Evol.* **32,** 803–805; 2017). If science isn't read, it doesn't exist.

One of the principal problems with writing a manuscript is that your individual voice is stamped out. Writers can be stigmatized by mentors, manuscript reviewers or journal editors if they use their own voice. Students tell me they are inspired to write, but worry that their adviser won't be supportive of creativity. It is a concern. We need to take a fresh look at the 'official style' — the dry, technical language that hasn't evolved in decades.

Author Helen Sword coined the phrase 'zombie nouns' to describe terms such as 'implementation' or 'application' that suck the lifeblood out of active verbs. We should engage readers' emotions and avoid formal, impersonal language. Still, there's a balance. Don't sensationalize the science. Once the paper has a clear message, I suggest that writers try some vivid language to help to tell the story. For example, I got some pushback on the title of one of my recent papers: 'Eight habitats, 38 threats, and 55 experts: Assessing ecological risk in a multi-use marine region'. But, ultimately, the editors let me keep it. There's probably less resistance out there than people might think.

Recently, after hearing me speak on this topic, a colleague mentioned that she had just rejected a review paper because she felt the style was too non-scientific. She admitted that she felt she had made the wrong decision and would try to reverse it.

BRETT MENSH

Create a logical framework

Scientific adviser, Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, Virginia; consultant, science communications.

Structure is paramount. If you don't get the structure right, you have no hope.

I co-wrote a paper (B. Mensh and K. Kording *PLoS Comput. Biol.* http://doi.org/ckqp; 2017) that lays out structural details for using a context–content–conclusion scheme to build a core concept. It is one of the most highly tweeted papers so far. In each paragraph, the first sentence defines the context, the body contains the new idea and the final sentence offers a conclusion. For the whole paper, the introduction sets the context, the results present the content and the discussion brings home the conclusion.

It's crucial to focus your paper on a single key message, which you communicate in the title. Everything in the paper should logically and structurally support that idea. It can be a delight to creatively bend the rules, but you need to know them first.

You have to guide the naive reader to the point at which they are ready to absorb what you did. As a writer, you need to detail the problem. I won't know why I should care about your experiment until you tell me why I should.

PETER GORSUCH

Prune that purple prose

Managing editor, Nature Research Editing Service, London; former plant biologist.

Writers must be careful about 'creativity'. It sounds good, but the purpose of a scientific paper is to convey information. That's it. Flourishes can be distracting. Figurative language can also bamboozle a non-native English speaker. My advice is to make the writing only as complex as it needs to be.

That said, there are any number of ways of writing a paper that are far from effective. One of the most important is omitting crucial information from the methods section. It's easy to do, especially in a complicated study, but missing information can make it difficult, if not impossible, to reproduce the study. That can mean the research is a dead end.

It's also important that the paper's claims are consistent with collected evidence. At the same time, authors should avoid being over-confident in their conclusions.

Editors and peer reviewers are looking for interesting results that are useful to the field. Without those, a paper might be rejected. Unfortunately, authors tend to struggle with the discussion section. They need to explain why the findings are interesting and how they affect a wider understanding of the topic. Authors should also reassess the existing literature and consider whether their findings open the door for future work. And, in making clear how robust their findings are, they must convince readers that they've considered alternative explanations.

STACY KONKIEL

Aim for a wide audience

Director of research and education at Altmetric, London, which scores research papers on the basis of their level of digital attention.

There have been no in-depth studies linking the quality of writing to a paper's impact, but a recent one (N. Di Girolamo and R. M. Reynders *J. Clin. Epidemiol.* **85,** 32–36; 2017) shows that articles with clear, succinct, declarative titles are more likely to get picked up by social media or the popular press.

Those findings tie in with my experience. My biggest piece of advice is to get to the point. Authors spend a lot of time setting up long-winded arguments to knock down possible objections before they actually state their case. Make your point clearly and concisely — if possible in non-specialist language, so that readers from other fields can quickly make sense of it.

If you write in a way that is accessible to non-specialists, you are not only opening yourself up to citations by experts in other fields, but you are also making your writing available to laypeople, which is especially important in the biomedical fields. My Altmetric colleague Amy Rees notes that she sees a trend towards academics being more deliberate and thoughtful in how they disseminate their work. For example, we see more scientists writing lay summaries in publications such as *The Conversation*, a media outlet through which academics share news and opinions. ■

INTERVIEWS BY VIRGINIA GEWIN

Interviews have been edited for clarity and length.

TURNING POINT
Speaking out

ERIKA PREUSS

Six months into the administration of US President Donald Trump, Joel Clement, former director of the Office of Policy Analysis at the US Department of the Interior, wrote an opinion piece in The Washington Post. *In it, he argued that his reassignment to an accounting post was intended to silence his work on climate-change adaptation among Alaska Native communities. Since leaving the agency last October, he has continued to advocate for the use of science when making policy decisions.*

TIMO KOHLER

You were hired under President Obama?
Yes. I came to Interior in 2011, after 8 years at a non-profit foundation in Seattle, Washington, that works on science-based conservation solutions. At Interior, which manages 75% of federal public lands, I built links with scientists, other federal agencies and political leaders. The most successful arena for doing that was climate adaptation and resilience — areas that Interior must address. No one was looking at the Arctic, where 60% of Interior-managed land is located, so I knew we needed to establish a climate-resilience plan for the region. In 2016, I co-chaired the Arctic Resilience Assessment (see go.nature.com/2gwfqrt) with Johan Rockström, executive director of the Stockholm Resilience Centre, which authored a plan for addressing crucial climate concerns that could reshape the Arctic, such as shifts in ice cover, extreme weather events and a collapse in fisheries.

Can you describe the Trump transition?
Nothing could have prepared me. I assured international colleagues that Arctic resilience would remain a priority. I was so wrong.

What prompted the opinion piece?
I'd worked on climate policy for seven years, and last July I was assigned to the Office of Natural Resources Revenue, which collects royalty cheques from oil and gas companies. I had been speaking out about the dangers of climate change for Alaska Native communities. In this new post, I could no longer do that. After speaking to a lawyer, it was clear that I had a legitimate whistleblower case.

Why did you resign?
After the opinion piece ran (see go.nature.com/2gwhdwh), I took a few days off. When I returned, I was floored by the support I got from Interior career folks, who showed me a stack of fan mail. I had no communications from high-level Interior staff. Once you are an official whistleblower, you have protections. When I realized that I could have more of an impact on this administration outside the agency, I wanted to go out with a clear message for Interior Secretary Ryan Zinke. On 6 October, I published my resignation letter, which called out poor leadership in the agency and the dangers of ignoring scientific expertise on climate change. It went viral.

Have other agency scientists contacted you for whistleblowing advice?
Dozens.

What did you learn from your experience?
If you think you have a valid whistleblower complaint, find a lawyer who is well versed in employment law, and determine whether it is valid. And scientific-integrity policies exist at every agency. I've heard of scientists being told they can't present their work at conferences, which is a violation of those policies.

What have you done since October?
I've never been busier. I have been talking to the media to hold this administration accountable. I've turned down job offers to get this message across.

Would you work for a federal agency again?
I would. I feel strongly about the role that a scientist can have in the federal government. At Interior alone, there are 70,000 employees. In the next five years, nearly half will be eligible for retirement. We could have a real generational change that brings new ideas and approaches into federal government. We'll need good, smart folks to get it right. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

COMPETITION

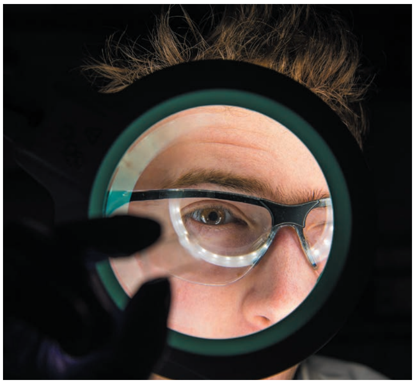
Science on camera

Do you like taking pictures of your work in the field — tracking polar bears, for example, scaling glaciers, examining fossils? Or of what you see daily in the laboratory, perhaps under a microscope or in a Petri dish?

If so, why not enter our second annual #ScientistAtWork photo competition? Here's how it works. Just send us your favourite picture (and any questions) to photocompetition@nature.com, or upload your image to Twitter or Instagram with the tag #ScientistAtWork. All entries (no more than one per person) must reach us by the end of March. The winning images will be chosen by an independent panel of *Nature* editors and art staff, who will be looking for pictures that tell an interesting story and reflect the variety of work that scientists do.

We will publish the five winning entries in the 26 April issue of *Nature*. As well as being highlighted in these pages, winners will receive a year's personal subscription to *Nature*, in print and online. (And eternal glory, of course.)

The stories that came with last years' images were as compelling as the photos themselves. After taking her winning entry, Kseniia Ashastina found herself staring down a gun barrel, having raised the suspicions of two locals who'd been digging up and selling the tusks of woolly mammoths (you can read how that turned out at go.nature.com/2gbob7w).



Microfluidic chip up close (a 2017 submission).

We're looking for images that highlight the working life of scientists in a dynamic, creative way. Heavily edited images will not be considered (see the terms and conditions on our dedicated blog at go.nature.com/scientistatwork), and photographs must be of high-enough quality (300 dpi at 220 millimetres wide) to appear in print.

Good luck, and we look forward to seeing your submissions. ■

LAVA CAKE FOR THE APOCALYPSE

A taste of history.

BY WENDY NIKEL

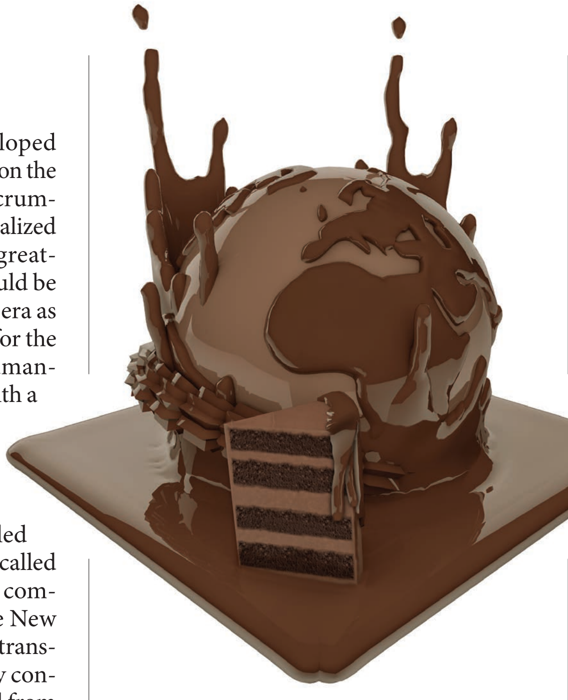


ILLUSTRATION BY JACEY

2 tablespoons wheat flour
Yes, Admiral, I know we have developed superior, genetically enriched starches on the New Worlds, but when I found this crumbling recipe card in the archives and realized it was written in your great-great-great-grandmother's hand, I thought it would be fitting to bake it as accurate to Earth-era as possible. As I was already scheduled for the reconnaissance mission back to humanity's birthplace, I thought I'd return with a little something extra to celebrate its impending destruction.

The flour wasn't too difficult to acquire, utilizing archived maps to pinpoint the area that used to be called the Wheat Belt. I don't know what it's called now; my search of the past century's communications between Earth and the New Worlds didn't provide any clues. The transmissions were infrequent and mainly consisted of lists of resources we required from Earth to enable our exploration and extend humanity's reach.

After our long flight across the stars and down through Earth's cloud-swept skies, Jabber and I set the Drifter down beside a rusted-out corpse of a tractor. I stared in awe at the amber waves rising up around us, still swaying in the breeze after all this time. Their seeds' texture on my tongue was peculiar, and I suddenly understood why we call things 'grainy'.

"Hurry up," Jabber called anxiously from the Drifter. "We've got some Earthers approaching from the north!"

Quickly, I stripped a handful of stalks and hurried back to the ship.

3 ounces chocolate
We headed south to find cacao trees on the plateau where my ancestors once lived.

After the Earthers' mutinous transmission — the one that started all this talk of restitution and repercussions and nuclear war — I spent some time alone in the archives, sifting through our collective history, trying without avail to uncover their motives.

I wondered: if I'd been planet-bound instead of sky-born, would I, too, be so backward-thinking and ignorant as to refuse our supply requests? Would I be so short-sighted as to deprive humanity of what we needed for life beyond Earth's orbit? I didn't know.

Maybe that's why I took your grandmother's notecard.

Sweat rolled from my forehead and my boots sank into the mud as I severed the cacao pods from their branches. Beyond the leaves, something moved: a barefoot child with hair the same shade as my own and eyes as dark as the void. He stared at me, a mirror to my past.

I held out a pod, pantomiming that he should take it, but instead, he turned and fled.

2 teaspoons sugar
I was still thinking of the child when we set the Drifter down within a field of green-leafed sugar cane. Over the hill crest lay a ruined city, and although Jabber advised me not to wander too close, now that I'd seen one Earther, I couldn't help my curiosity.

Far below, the people worked, hunched over, their arms moving in rapid rhythm. Through the heat of the day, they performed the tasks of the New Worlds' farmbots with their own hands while the machines of their city crumbled around them.

The sugar cane felt rough in my hand. Had they really produced all this with such primitive methods? Where were the machines, the technology, that had once elevated our ancestors to the skies?

I recalled our last transmission — requesting materials that the Earthers claimed were now impossible to procure, impossible to deliver to the rendezvous point — and considered for once that they might have been right.

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

We might have been wrong.

1 egg
As Jabber flew, I scanned the ground, searching now for what wasn't there. In a thick clump of trees loomed the shadow of a forgotten factory, its smokestacks twined with lush vines. Beneath the river's surface was the outline of a bus, draped with olive-green algae. A field of stones slowly resolved into the shape of a town, abandoned streets delineated with darker grey.

On and on, we flew past the shadows of our civilization.

We landed near a farm with a chicken coop beside it, amid a ramshackle cluster of houses. I ducked inside the tiny enclosure to pocket a single, warm egg.

"Who are you?" A woman wearing an apron suddenly blocked my path, leading a bored, brown cow. Jabber would kill me if he knew I'd let them surprise me.

I raised my hands and said the only thing that came to mind: "Please, ma'am. Do you have any butter?"

1 teaspoon butter
Jabber didn't bother asking where I'd got the hand-painted bowl with the generous dollop of butter cradled inside. I think he was just glad to leave. But all the while, as the Drifter sped across the Galaxy and Earth grew smaller in the window, I couldn't stop thinking of the empty cities, the apron-clad milkmaid, and the child with the void in his eyes.

The cake fell in the middle and the edges are too crisp, but I want you to have it regardless. I want you to taste for yourself your great-great-great-grandmother's world — sweet and bitter and complex.

Taste a world worth saving. ■

Wendy Nikel *is a speculative-fiction author with a degree in elementary education, a fondness for road trips and a terrible habit of forgetting where she's left her cup of tea. Her short fiction has been published by* Fantastic Stories of the Imagination, Daily Science Fiction, Nature *and elsewhere. Her time-travel novella,* The Continuum, *was published by World Weaver Press in January 2018. For more info, visit wendynikel.com.*